

THE BUCHAREST UNIVERSITY OF ECONOMIC STUDIES

# DATABASE SYSTEMS JOURNAL

**Vol. XVI/2025**

## LISTED IN

RePEc, EBSCO, DOAJ, Open J-Gate,  
Cabell's Directories of Publishing Opportunities,  
Index Copernicus, Google Scholar,  
Directory of Science, Cite Factor,  
Electronic Journals Library

BIG DATA

NoSQL

DATA SCIENCE

MACHINE LEARNING

BUSINESS INTELLIGENCE

CLOUD COMPUTING

DATA MINING

DATA WAREHOUSES

DATABASES

**ISSN: 2069 – 3230**  
dbjournal.ro

## Database Systems Journal BOARD

### Director

Prof. Ion Lungu, PhD, University of Economic Studies, Bucharest, Romania

### Editors-in-Chief

Prof. Adela Bara, PhD, University of Economic Studies, Bucharest, Romania

Conf. Iuliana Botha, PhD, University of Economic Studies, Bucharest, Romania

### Editors

Conf. Anca Andreescu, PhD, University of Economic Studies, Bucharest, Romania

Conf. Anda Belciu, PhD, University of Economic Studies, Bucharest, Romania

Prof. Ramona Bologa, PhD, University of Economic Studies, Bucharest, Romania

Prof. Vlad Diaconița, PhD, University of Economic Studies, Bucharest, Romania

Conf. Alexandra Florea, PhD, University of Economic Studies, Bucharest, Romania

Prof. Adina Uța, PhD, University of Economic Studies, Bucharest, Romania

### Editorial Board

Prof. Ioan Andone, PhD, A.I.Cuza University, Iasi, Romania

Prof. Emil Burtescu, PhD, University of Pitesti, Pitesti, Romania

Joshua Cooper, PhD, Hildebrand Technology Ltd., UK

Prof. Marian Dardala, PhD, University of Economic Studies, Bucharest, Romania

Prof. Dorel Dusmanescu, PhD, Petrol and Gas University, Ploiesti, Romania

Prof. Marin Fotache, PhD, A.I.Cuza University, Iasi, Romania

Dan Garlasu, PhD, Oracle Romania

Prof. Marius Guran, PhD, University Politehnica of Bucharest, Bucharest, Romania

Lect. Ticiano Costa Jordão, PhD-C, University of Pardubice, Pardubice, Czech Republic

Prof. Brijender Kahanwal, PhD, Galaxy Global Imperial Technical Campus, Ambala, India

Prof. Dimitri Konstantas, PhD, University of Geneva, Geneva, Switzerland

Prof. Hitesh Kumar Sharma, PhD, University of Petroleum and Energy Studies, India

Prof. Marinela Mircea, PhD, University of Economic Studies, Bucharest, Romania

Prof. Mihaela I.Muntean, PhD, West University, Timisoara, Romania

Prof. Stefan Nitchi, PhD, Babes-Bolyai University, Cluj-Napoca, Romania

Prof. Corina Paraschiv, PhD, University of Paris Descartes, Paris, France

Davian Popescu, PhD, Milan, Italy

Prof. Gheorghe Sabau, PhD, University of Economic Studies, Bucharest, Romania

Prof. Nazaraf Shah, PhD, Coventry University, Coventry, UK

Prof. Ion Smeureanu, PhD, University of Economic Studies, Bucharest, Romania

Prof. Traian Surcel, PhD, University of Economic Studies, Bucharest, Romania

Prof. Ilie Tamas, PhD, University of Economic Studies, Bucharest, Romania

Silviu Teodoru, PhD, Oracle Romania

Prof. Dumitru Todoroi, PhD, Academy of Economic Studies, Chisinau, Republic of Moldova

Prof. Manole Velicanu, PhD, University of Economic Studies, Bucharest, Romania

Prof. Robert Wrembel, PhD, University of Technology, Poznan, Poland

### Contact

Calea Dorobanților, no. 15-17, room 2017, Bucharest, Romania

Web: <https://dbjournal.ro/>

E-mail: [editordbjournal@gmail.com](mailto:editordbjournal@gmail.com)

## CONTENTS

<b>The Use of Communication Platforms in Military Operations: Enhancing Strategic and Tactical Effectiveness .....</b>	<b>1</b>
Rexhep MUSTAFOVSKI	
<b>Query Completion for Small-Scale Distributed Databases in PostgreSQL and MongoDB .....</b>	<b>11</b>
Marin FOTACHE, Cătălina BADEA, Marius-Iulian CLUCI, Codrin-Ştefan EŞANU	
<b>An Overview of Big Data and NoSQL in the Video Game Industry .....</b>	<b>28</b>
Cristiana COSTAN	
<b>Evaluating Deep Learning and Machine Learning Models in Federated Learning for Credit Card Fraud Detection. A comparative study .....</b>	<b>37</b>
Şener ALI	
<b>The Advantage of NoSQL Databases over SQL Databases .....</b>	<b>45</b>
Mario-Tudor CHIRIAC	
<b>House Market Prediction Using Machine Learning .....</b>	<b>55</b>
Nicuşor-Andrei ANDREI	
<b>Detection of Fake News Using Deep Learning and Machine Learning .....</b>	<b>65</b>
Gabriela CHIRIAC, Ada Maria CATINA	
<b>Big Data and Business Intelligence: Synergies for Leveraging Competitive Advantage .....</b>	<b>82</b>
Alexandru-Dan TURIAC, Mihai-Paul UNGUREANU, Mihnea-Cristian TACHE	
<b>The Role of Data Science and Machine Learning in Strategic Decision Making .....</b>	<b>98</b>
Elena-Luiza JALEA, Marian-Robert VLĂDUŢĂ	
<b>Comparative Aspects of E-learning Platforms in the Field of Foreign Languages .....</b>	<b>114</b>
Georgiana VOICA	
<b>Comparation of the Performance of SQL and NoSQL Databases in Modern Business Applications .....</b>	<b>126</b>
Aniela BORCAN, Diana CAPTARI, Emanuela-Cristina CARP	
<b>A Taxonomy for Learning Cybersecurity with Artificial Intelligence .....</b>	<b>143</b>
Mihai-Nicolae DULGHERU, Andrei-Cătălin NICA, Cătălin-Alexandru TRANDAFIRIDIS	

# The Use of Communication Platforms in Military Operations: Enhancing Strategic and Tactical Effectiveness

Rexhep MUSTAFOVSKI  
Ss. Cyril and Methodius University  
Faculty of Electrical Engineering and Information Technologies  
Skopje, Republic of North Macedonia  
[rexhepmustafovski@gmail.com](mailto:rexhepmustafovski@gmail.com)

**Abstract:** *The way communication platforms are used in military operations has changed a lot over the years. They're now essential for mission success, quick decision-making, and maintaining strategic advantages. This paper dives into the modern communication tools that are making waves in military settings, with a spotlight on networked communication, signal support, cybersecurity strategies, and how different forces work together in joint and multinational missions. It also tackles some of the major hurdles, like congestion in the electromagnetic spectrum, cyber threats, and the need for secure data transmission in challenging environments. By pulling insights from FM 6-02: Signal Support to Operations and the latest scientific research, this paper highlights recent advancements in military communication tech and how they boost operational effectiveness. Additionally, the research looks ahead at future possibilities, such as AI-driven communication platforms, quantum encryption, and cutting-edge satellite networks for defense purposes. This paper's primary contribution is the development of a structured, AI-enabled communication workflow that integrates quantum-safe encryption, blockchain authentication, and satellite-based coordination to improve decision-making and resilience in multi-domain operations.*

**Keywords:** *Military communication, secure networks, battlefield connectivity, radio systems, cybersecurity, signal support, tactical communications, AI in defense*

## 1 Introduction

The effectiveness of military operations depends heavily on reliable, secure, and adaptable communication platforms. As modern warfare becomes increasingly network-centric, the ability to transmit real-time information across multiple operational domains—land, air, sea, space, and cyberspace—has become a strategic necessity. Military communication systems have evolved to integrate advanced radio networks, satellite-based communication, and encrypted data transmission, ensuring seamless coordination between deployed forces and command centers [1], [3], [5]. The increasing complexity of multinational operations, joint task forces, and asymmetric warfare has further emphasized the need for highly secure and interoperable communication infrastructures that can operate even in contested environments [6]–[8].

A key aspect of modern military communication is signal support, which provides robust network connectivity and ensures continuous information flow despite cyber threats, electronic warfare (EW) attacks, and environmental disruptions. Military doctrines emphasize the importance of agility, redundancy, and survivability in communication networks, ensuring that forces remain operational even when primary channels are compromised [9], [11]. The U.S. Army's Field Manual 6-02 (FM 6-02) outlines fundamental principles of signal support, including interoperability, network resilience, and cybersecurity strategies essential for mission success [17]. These principles are crucial in large-scale combat operations, where the ability to maintain command and control (C2) can determine operational success or failure [12], [14]. The transition from analog to digital military communication platforms has introduced a range of advantages, such as higher data

transfer rates, AI-assisted network management, and automated encryption protocols [15]–[17]. The introduction of software-defined radios (SDRs), wideband satellite networks, and AI-driven cyber defense mechanisms has further enhanced battlefield communication, allowing for real-time situational awareness and faster decision-making [4], [10]. However, despite these advancements, challenges remain, particularly regarding electromagnetic spectrum congestion, jamming threats, and cyberattacks from adversarial forces [7], [9].

Cybersecurity remains one of the most critical concerns in military communication platforms. Cyber warfare, espionage, and digital sabotage pose a significant risk to military networks, requiring advanced encryption techniques, quantum-safe cryptographic methods, and AI-based intrusion detection systems to mitigate vulnerabilities [6], [8]. The Department of Defense Information Network (DODIN) plays a crucial role in securing data transmission, integrating multi-layer encryption protocols, blockchain-based authentication, and adaptive firewall mechanisms [13], [16]. The role of artificial intelligence and machine learning in automating cybersecurity responses and threat detection is expected to further revolutionize military communication strategies [14]–[17].

This paper explores the current state of military communication platforms, their integration with modern defense technologies, and the challenges posed by electronic warfare and cybersecurity threats. The study also discusses future trends in military communication, including quantum encryption, AI-assisted battlefield networking, and low-Earth orbit (LEO) satellite communication systems [2], [5], [12]. By analyzing these technological advancements and their impact on military operations, this research aims to

provide valuable insights into the future of secure and resilient communication platforms in defense applications [1], [4], [8].

The main scientific contribution of this study lies in its proposal of an adaptive workflow for next-generation military communication platforms. By integrating AI-driven automation, blockchain security, and low-latency satellite connectivity, the paper introduces a novel architecture that significantly enhances tactical coordination, cyber resilience, and real-time situational awareness in hostile environments.

## **2 Literature Review**

The study of military communication platforms has evolved significantly in recent years, with various technological advancements, operational challenges, and security concerns shaping the field. Several scholarly works and military doctrines have explored the integration of digital communication networks, cybersecurity measures, and tactical communication strategies in modern warfare [1]–[3]. The transition from analog to digital systems, the emergence of AI-driven networking, and the adoption of secure satellite communications have played a crucial role in improving the efficiency of military communication platforms [4], [5].

One of the key components of military communication is interoperability, which ensures seamless information exchange between different branches of the armed forces and multinational coalitions [6], [7]. The U.S. Army's Field Manual FM 6-02 emphasizes the importance of joint operations, where different communication networks must operate efficiently under a unified system [17]. Research highlights the role of secure radio networks, software-defined radios (SDRs), and advanced satellite communication systems in achieving this goal [8], [9]. Table 1 provides an overview of traditional vs. modern military communication platforms, showing their differences in terms of data transmission, security, and adaptability.

Table 1. Comparison of Traditional and Modern Military Communication Platforms

Feature	Traditional Communication	Modern Digital Communication
Transmission Medium	Analog radio signals	Encrypted digital networks
Data Rate	Low	High-speed broadband
Security	Vulnerable to interception	End-to-end encryption & AI-based security
Interoperability	Limited	Fully integrated with joint/multinational operations
Adaptability	Fixed-frequency systems	Software-defined, frequency-agile systems
Reliability in Combat	Prone to jamming	Resistant to EW & cyber threats

While prior studies have discussed individual elements such as AI, SDRs, or satellite systems, this paper uniquely proposes an integrated framework combining these technologies with blockchain and quantum encryption. The proposed workflow advances beyond existing architecture by offering real-time threat adaptation, predictive decision support, and unified cross-domain operations, which are not concurrently addressed in existing literature.

2.1 Advancements in Military Communication Platforms

The integration of AI and machine learning (ML) in military communication networks has allowed automated threat detection, real-time data processing, and network self-healing capabilities [10], [11]. AI-based systems can predict cyber

threats, manage bandwidth distribution, and optimize data transmission in congested environments, significantly enhancing operational efficiency [12], [13]. Satellite communication (SATCOM) plays a pivotal role in long-range military operations, providing beyond-line-of-sight (BLOS) connectivity, global reconnaissance, and secure battlefield networking [14], [15]. Low-Earth Orbit (LEO) satellites have gained popularity due to their low latency, high-speed data transmission, and enhanced resilience against jamming [16], [17]. These systems complement traditional geostationary satellites (GEO), ensuring continuous coverage in remote and contested environments. Table 2 highlights the differences between various satellite communication technologies used in military operations.

Table 2. Comparison of Military Satellite Communication Technologies

Feature	Geostationary Satellites (GEO)	Medium Earth Orbit (MEO)	Low Earth Orbit (LEO)
Altitude	~35,786 km	~5,000–20,000 km	~500–1,500 km
Latency	High (~500ms)	Moderate (~200ms)	Low (~50ms)
Coverage	Global but with delay	Regional	High-speed global coverage
Jamming Resistance	Moderate	High	Very High
Data Transfer Speed	Limited	Moderate	High-speed broadband
Deployment Cost	High	Moderate	Lower than GEO/MEO

## 2.2 Cybersecurity Challenges in Military Communication

As digital transformation continues in military operations, cybersecurity remains a critical challenge. Hostile actors, state-sponsored cyber units, and non-state entities are continuously attempting to exploit vulnerabilities in military networks [5], [6]. Recent cyber-attacks on defense infrastructures highlight the necessity for multi-layered encryption, real-time intrusion detection, and AI-driven anomaly detection systems [7]–[9].

Cyber threats can disrupt battlefield communications, intercept classified intelligence, or manipulate strategic data, leading to severe consequences on national security. The FM 6-02 doctrine outlines cyber protection measures, including encryption protocols, threat monitoring, and secure data transmission techniques to mitigate these risks [17].

To address these cybersecurity concerns, military forces are investing in post-quantum encryption, blockchain authentication, and AI-driven cybersecurity tools [10], [11]. Automated network defense mechanisms, combined with zero-trust security models, are expected to play a crucial role in securing military communication networks [12], [13].

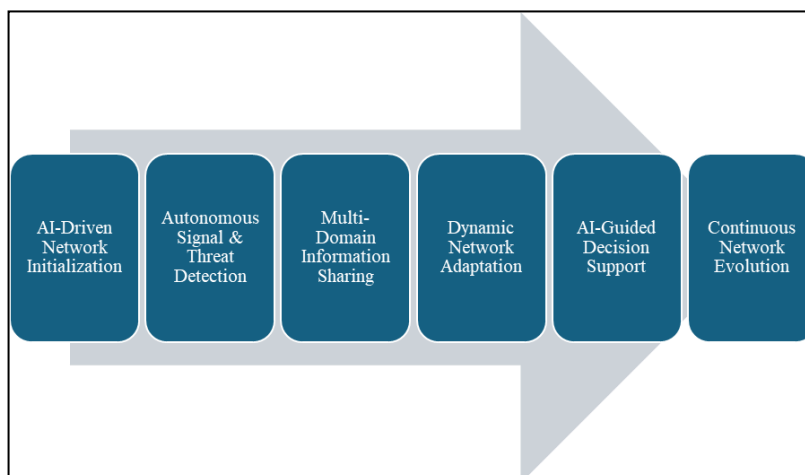
## 2.3 Future Trends in Military Communication Platforms

The future of military communication is being shaped by AI-driven automation, quantum-safe encryption, and next-generation tactical communication platforms. Some of the emerging trends include:

- **AI-Powered Autonomous Networks:** AI-driven systems will automate network management, optimize signal transmission, and predict potential threats [14].
- **Quantum Communication:** The adoption of quantum key distribution (QKD) and quantum-safe encryption will enhance data security in military operations [15], [16].
- **Integrated IoT-based Defense Systems:** The Internet of Battlefield Things (IoBT) will enable real-time sensory integration, predictive analytics, and automated reconnaissance [17].
- **Augmented Reality (AR) Communication:** The use of AR and holographic battlefield interfaces will improve situational awareness and mission coordination [9], [10].

## 3 A New Workflow for the Next Generation of Military Communication Systems

The proposed workflow consists of six key stages, ensuring secure, adaptive, and mission-driven communication that enhances interoperability among military branches, joint task forces, and multinational coalitions.



**Fig. 1.** Workflow for the Next Generation of Military Communication Systems

• **Workflow Steps and Explanation:**

1. **AI-Driven Network Initialization & Configuration**

- The network automatically establishes a secure and encrypted connection across all deployed units.
- AI-based network optimization identifies the best communication pathways to minimize latency and bandwidth congestion.
- Quantum encryption secures data transmission from the start to prevent cyberattacks.

2. **Autonomous Tactical Signal Management & Threat Detection**

- AI-powered systems continuously monitor radio frequency (RF) environments for jamming attempts or cyber threats.
- Smart spectrum management automatically adjusts frequency usage, ensuring optimal connectivity in contested environments.
- Blockchain-based authentication ensures that only authorized personnel can access sensitive communications.

3. **Real-Time Multi-Domain Information Sharing**

- Edge computing and IoT sensors collect, analyze, and distribute battlefield intelligence instantly.
- AI-based speech recognition converts voice commands into encrypted digital messages for secure, instant dissemination.
- Unified Command Network (UCN) ensures seamless

communication between land, air, sea, space, and cyber forces.

4. **Dynamic Network Adaptation in Combat Environments**

- Low-Earth Orbit (LEO) satellite networks provide real-time battlefield coverage, reducing reliance on vulnerable ground infrastructure.
- SDR-based radio networks dynamically switch frequencies to avoid electronic warfare (EW) interference.
- Cloud-based situational awareness platforms allow commanders to access real-time intelligence anywhere.

5. **AI-Guided Decision Support & Predictive Analytics**

- AI-powered battlefield assistants analyze incoming data and recommend the best tactical responses based on real-time intelligence.
- Predictive cybersecurity models anticipate cyber threats and apply preemptive countermeasures.
- Automated logistical coordination ensures supply chains remain intact and uninterrupted.

6. **Continuous Network Evolution & Learning**

- Self-healing networks use machine learning (ML) to identify weak points and automatically reinforce connectivity.
- AI-based after-action reviews (AARs) analyze communication effectiveness and optimize future missions.
- Secure post-mission data storage ensures historical operational records remain classified and protected.



**Table 3.** Workflow for Next-Generation Military Communication

Step	Process	Technology Used	Impact on Military Operations
<b>1. AI-Driven Network Initialization</b>	Automatic secure connections & encrypted channels	AI-driven routing, quantum encryption	Faster & safer mission readiness
<b>2. Autonomous Signal &amp; Threat Detection</b>	Real-time RF scanning & adaptive spectrum use	AI-assisted EW defense, blockchain security	Enhanced jamming resistance & secure access control
<b>3. Multi-Domain Information Sharing</b>	Seamless cross-platform communication	IoT-based intelligence, AI speech recognition	Faster battlefield awareness
<b>4. Dynamic Network Adaptation</b>	Resilient satellite & SDR-based networks	LEO satellites, cloud-based ops centers	Reliable comms in contested environments
<b>5. AI-Guided Decision Support</b>	AI-driven battlefield analysis & predictive logistics	AI-powered analytics threat modeling	Improved decision-making speed
<b>6. Continuous Network Evolution</b>	Self-healing networks & adaptive AI	ML-based security, after-action data analysis	Long-term resilience & future optimization

The proposed next-generation military communication workflow will revolutionize military operations by providing:

**1. Uninterrupted Communication in High-Risk Environments**

- AI-powered adaptive frequency shifting and SDR ensure continuous connectivity even under electronic warfare conditions.
- Decentralized blockchain authentication prevents data spoofing or interference.
- LEO satellites ensure resilient BLOS (Beyond-Line-of-Sight) communication in remote areas.

**2. Improved Decision-Making & Tactical Coordination**

- AI-driven command assistants analyze real-time battlefield data and offer rapid, informed decision recommendations.
- Enhanced situational awareness allows commanders to adapt quickly based on automated intelligence reports.
- Interoperable networks enable seamless collaboration among land, air, sea, and cyber units.

**3. Faster Deployment & Readiness**

- AI-driven network automation reduces deployment times by instantly configuring communication systems.
- Autonomous security systems ensure data protection from cyberattacks before missions begin.
- Predictive analytics assist in early detection of potential threats, ensuring forces are prepared.

**4. Advanced Cybersecurity Protection Against Digital Threats**

- AI-driven cybersecurity models predict malware, phishing, and cyber-intrusion attempts before they happen.
- Quantum-safe encryption ensures that classified military data remains unbreakable.
- Dynamic firewall systems prevent adversaries from exploiting communication vulnerabilities.

**5. Greater Interoperability Between Allied Forces**

- The Unified Command Network (UCN) ensures that all allied forces, regardless of branch or nation, can

- communicate without compatibility issues.
- Standardized AI-assisted voice translation systems allow seamless communication between different military forces in multinational operations.
- The integration of cloud-based command centers enables real-time data access from global military installations.

**Table 4.** Impact of Next-Generation Military Communication on Operational Success

Key Advantage	Traditional Communication	Next-Generation Communication	Impact on Operations
Speed of Deployment	Manual setup & configuration	AI-automated network setup	Reduced setup time & faster mission readiness
Reliability in Combat	Prone to jamming	Adaptive SDRs & LEO satellites	Continuous connection in high-risk areas
Cybersecurity	Encryption only	AI-driven security & quantum-safe encryption	Stronger data protection & cyber resilience
Interoperability	Limited to branch-specific networks	Unified Command Network (UCN)	Seamless allied coordination
Threat Response	Manual security checks	AI-assisted real-time anomaly detection	Preemptive cyber defense measures

**4 Challenges in Implementing the Proposed Next-Generation Military Communication Platform**

While the proposed AI-driven, quantum-encrypted, and multi-domain military communication platform offers significant advantages in security, interoperability, and efficiency, its implementation presents several challenges. Military environments demand high resilience, rapid adaptability, and uncompromising security, making the integration of advanced technologies a complex and resource-intensive endeavour. Below are the key challenges that must be addressed for successful deployment.

**a. Cybersecurity Threats and Quantum-Resistant Encryption**

One of the most significant challenges in implementing this next-generation communication platform is ensuring robust cybersecurity. Military networks are prime targets for cyber warfare, espionage, and digital sabotage by state-sponsored actors and sophisticated

adversaries. The adoption of quantum encryption enhances security, but the transition from traditional encryption methods poses integration challenges, high computational costs, and compatibility issues with legacy systems. Additionally, AI-powered cybersecurity tools must be rigorously tested to avoid vulnerabilities that adversaries could exploit.

**b. Electromagnetic Spectrum Management and Jamming Resistance**

Modern warfare relies heavily on the electromagnetic spectrum for secure radio, satellite, and networked communication. However, spectrum congestion, frequency interference, and electronic warfare (EW) attacks pose severe threats. Adaptive frequency hopping, AI-driven spectrum allocation, and SDR (Software-Defined Radios) help mitigate these risks, but their deployment requires extensive testing in real-world combat scenarios. Additionally, adversaries constantly develop advanced jamming and spoofing techniques, making continuous adaptation and countermeasure development essential.

**c. Interoperability with Legacy and Allied Systems**

Military forces operate with a mix of old and new communication technologies, making interoperability a major concern. Many allied forces and joint coalitions use different encryption standards, radio frequencies, and data-sharing protocols, requiring a universal framework to ensure seamless communication. The proposed Unified Command Network (UCN) can enhance cross-force connectivity, but integrating diverse national defense systems into a single secure framework is a technically and politically complex challenge.

**d. AI and Automation Risks in Battlefield Decision-Making**

The integration of AI-driven decision support systems and automated battlefield analytics enhances efficiency but also introduces risks related to machine bias, decision latency, and potential adversarial manipulation. AI must interpret real-time battlefield data with high accuracy, but errors in object classification, speech recognition, or enemy movement prediction can lead to tactical failures. Additionally, AI's reliance on massive data sets raises concerns about data security, ethical considerations, and the risk of enemy AI-based countermeasures.

**e. Infrastructure Deployment in Remote and Hostile Environments**

Deploying a resilient, battlefield-ready communication infrastructure requires secure, mobile, and easily maintainable networking equipment. Remote operations—such as desert, jungle, or arctic warfare—face logistical challenges in establishing LEO satellite links, maintaining power sources, and protecting hardware from environmental stressors. Military-grade energy-efficient, solar-powered, and autonomous relay nodes can help mitigate these issues, but

ensuring long-term network uptime remains a challenge.

**f. High Costs and Budget Constraints**

The development, testing, and large-scale implementation of next-generation military communication networks require substantial investment. Quantum-safe encryption, AI-driven cybersecurity, and satellite-based networking involve expensive research, hardware procurement, and system training costs. Many military organizations face budget constraints, long procurement cycles, and policy debates that could delay adoption and full-scale deployment.

**g. Resistance to Change and Human Adaptation**

The shift to AI-enhanced, real-time, and automated communication platforms may face resistance from military personnel accustomed to traditional systems. Soldiers and commanders require extensive training to operate new encrypted networks, AI-driven interfaces, and cloud-based data-sharing platforms. Without proper user adaptation, trust, and operational readiness, even the most advanced system could fail under real-world combat pressure.

**5 Conclusion**

The rapid advancement of military communication platforms has redefined modern warfare, enabling faster decision-making, improved battlefield coordination, and enhanced cybersecurity. As conflicts become more technologically driven, military forces must adopt AI-driven automation, quantum-safe encryption, and real-time satellite-based connectivity to maintain operational superiority. The transition from traditional analog and single-frequency radio systems to adaptive, software-defined networks and AI-enhanced communication infrastructures ensures that military units remain connected even in contested, high-risk environments. Interoperability remains a cornerstone of future military communication, enabling seamless integration among different

branches, joint forces, and multinational coalitions. The Unified Command Network (UCN), enhanced by AI-assisted real-time speech translation and blockchain authentication, will allow allied forces to share intelligence, coordinate logistics, and execute joint operations without compatibility issues. Additionally, the deployment of LEO satellite networks will significantly reduce reliance on vulnerable ground-based infrastructure, ensuring uninterrupted connectivity and beyond-line-of-sight (BLOS) communication in remote battlefields. Cybersecurity is another crucial factor in military communication, as digital threats continue to evolve. AI-powered threat detection, self-healing networks, and quantum-resistant encryption will provide unparalleled protection against cyber espionage, electronic warfare, and digital sabotage. The integration of predictive analytics and automated cyber-defense mechanisms ensures preemptive threat mitigation before hostile actors can exploit vulnerabilities.

The next generation of military communication will be more adaptive, intelligent, and secure, offering greater operational flexibility and resilience. By integrating AI, autonomous network management, and multi-domain interoperability, military forces can achieve superior situational awareness, faster response times, and enhanced combat effectiveness. These advancements will not only strengthen battlefield capabilities but also ensure long-term strategic dominance in modern and future conflicts. Military success will increasingly depend on who can communicate faster, more securely, and more efficiently in the evolving digital battlespace.

## References

- [1] Abdelzaher, Tarek / Wigness, Mike / Russell, Sean / Swami, Ananthram: *Internet of Battlefield Things: Challenges, Opportunities, and Emerging Directions*. IoT for Defense and National Security, 2023, pp. 5–22.
- [2] Akbar, Ridho S / Kholid, Fadhil / Kasiyanto, Kharis / Widiatmoko, Donny / Achmad, Agus: *Design of Fuel Monitoring Application for Reservoir Tanks in Army Fuel Supply Point on Military Logistics Corps Based on Internet of Things*. International Journal of Engineering and Computer Science Applications (IJECSA), 2024, pp. 19–32.
- [3] Alkanjr, Bilal / Mahgoub, Ibrahim: *A Novel Deception-Based Scheme to Secure the Location Information for IoBT Entities*. IEEE Access, 2023, pp. 15–554.
- [4] Azar, Jad / Makhoul, Amal / Barhamgi, Mounira / Couturier, Raphael: *An Energy Efficient IoT Data Compression Approach for Edge Machine Learning*. Future Generation Computer Systems, 2023, pp. 168–175.
- [5] Butun, Ismail / Mahgoub, Ibrahim: *Expandable Mix-Zones as a Deception Technique for Providing Location Privacy on Internet-of-Battlefield Things (IoBT) Deployments*. IEEE Access, 2024.
- [6] Doku, Reginald / Rawat, Danda B / Garuba, Muritala / Njilla, Laurent: *Fusion of Named Data Networking and Blockchain for Resilient Internet-of-Battlefield-Things*. IEEE 17th Annual Consumer Communications & Networking Conference (CCNC), 2023, pp. 1–6.
- [7] Farooq, M. Junaid / Zhu, Quanyan: *Secure and Reconfigurable Multi-Layer Network Design for Critical Information Dissemination in the Internet of Battlefield Things (IoBT)*. IEEE Transactions on Wireless Communications, 2023, pp. 2618–2632.
- [8] Feng, Yujie / Li, Ming / Zeng, Chen / Liu, Hu: *Robustness of Internet of Battlefield Things (IoBT): A Directed Network Perspective*. Entropy, 2023, pp. 1166.

- [9] Heidari, Alireza / Jamali, Jabraeil: *Internet of Things Intrusion Detection Systems: A Comprehensive Review and Future Directions*. Cluster Computing, 2023, pp. 3753–3780.
- [10] Joshi, Suraj / Thakar, Anshuman / Patel, Chirag: *Applications of Machine Learning and Deep Learning in Securing Internet of Battlefield Things: A Futuristic Perspective*. 2023 10th International Conference on Computing for Sustainable Global Development (INDIACom), IEEE, 2023, pp. 333–338.
- [11] Karim, Hossain / Rawat, Danda B: *Evaluating Machine Learning Classifiers for Data Sharing in Internet of Battlefield Things*. IEEE Symposium Series on Computational Intelligence (SSCI), 2023, pp. 1–7.
- [12] Kufakunesu, Rachel / Myburgh, Herman / De Freitas, Allan: *The Internet of Battle Things: A Survey on Communication Challenges and Recent Solutions*. Discover Internet of Things, 2024, pp. 3–19.
- [13] Kunatsa, Tendai / Myburgh, Herman C / De Freitas, Allan: *Optimal Power Flow Management for a Solar PV-Powered Soldier-Level Pico-Grid*. Energies, 2024, pp. 459.
- [14] Liu, Dandan / Abdelzaher, Tarek / Wang, Tianyu / Hu, Yibo / Li, Jian / Liu, Shiyang / Caesar, Matthew / Kalasapura, Deepak / Bhattacharyya, Jay / Srouf, Nabil: *IoBT-OS: Optimizing the Sensing-to-Decision Loop for the Internet of Battlefield Things*. IEEE International Conference on Computer Communications and Networks (ICCCN), 2023, pp. 1–10.
- [15] Masuduzzaman, Mohammad / Rahim, Tariqul / Islam, Asif / Shin, Sun Yoon: *UAV-Employed Intelligent Approach to Identify Injured Soldier on Blockchain-Integrated Internet of Battlefield Things*. IEEE Transactions on Network and Service Management, 2024.
- [16] Rutravigneshwaran, Prabu / Anitha, Gunasekaran / Prathapchandran, Krishnakumar: *Trust-Based Support Vector Regressive (TSVR) Security Mechanism to Identify Malicious Nodes in the Internet of Battlefield Things (IoBT)*. International Journal of System Assurance Engineering and Management, 2024, pp. 287–299.
- [17] Singh, Rakesh Kumar / Mishra, Satyendra: *TinyML Meets IoBT Against Sensor Hacking*. The Network and Distributed System Security (NDSS) Symposium, Workshop on Security and Privacy in Standardized IoT (SDIoTSec), 2024, pp. 1–9.



**Rexhep Mustafovski, MSc**, is a Teaching and Research Assistant at the Military Academy "General Mihailo Apostolski" - Skopje, Department for Cybersecurity and Digital Forensics. He holds a Master's degree from the Faculty of Electrical Engineering and Information Technologies, University "Ss. Cyril and Methodius" - Skopje. His research interests include radar systems, IoT security, object detection technologies, and integrated control and monitoring systems. He has authored and co-authored more than 20 scientific and professional papers published in international conferences and

journals, including several indexed in high-impact factor journals. His work contributes to advancements in next-generation radar systems, cybersecurity, and military applications of digital technologies.

## Query Completion for Small-Scale Distributed Databases in PostgreSQL and MongoDB

Marin FOTACHE<sup>1</sup>, Cătălina BADEA<sup>2</sup>, Marius-Iulian CLUCI<sup>3</sup>, Codrin-Ștefan EȘANU<sup>4</sup>

<sup>1</sup>Dept. of Accounting, Information Systems and Statistics

Alexandru Ioan Cuza University of Iași, Faculty of Economics and Business Administration

<sup>2</sup>Alexandru Ioan Cuza University of Iași

<sup>3</sup> Alexandru Ioan Cuza University of Iași, S.C. HVM SRL

<sup>4</sup> Alexandru Ioan Cuza University of Iași, Cegeka Romania  
Iași, Romania

[fotache@uaic.ro](mailto:fotache@uaic.ro), [catalina.badea@student.uaic.ro](mailto:catalina.badea@student.uaic.ro), [marius.cluci@uaic.ro](mailto:marius.cluci@uaic.ro),  
[codrin.stefan@feaa.uaic.ro](mailto:codrin.stefan@feaa.uaic.ro)

*Relational/SQL and document/JSON data stores are competing but also complementary technologies in the OLAP (On-Line Analytical Processing) systems. Whereas the traditional approaches for performance comparison use the duration of queries performing similar tasks, in this paper we compare the performance of two distributed setups deployed on PostgreSQL/Citus and MongoDB by focusing only on the query's successful completion within a 10-minute timeout. The TPC-H benchmark database was converted into a denormalized JSON schema in MongoDB. An initial set of 296 SQL queries was devised for execution in PostgreSQL/Citus and then mapped for execution in MongoDB using Aggregation Framework (AF). Query execution success within a 10-minute timeout was collected for both PostgreSQL and MongoDB in six scenarios defined by two small-scale data factors (0.01 and 0.1 GB) and three different node counts (3, 6, and 9) for data distribution and processing. The relationships between the query completion and the query parameters were assessed with statistical tests and a series of machine learning techniques.*

**Keywords:** PostgreSQL, Citus, MongoDB, SQL, Aggregation Framework, OLAP performance comparison.

### 1 Introduction

Current data architectures incorporate various technologies for data storage and processing, such as relational/SQL and NoSQL data servers [1][2][3]. In this paper, we analyse and compare the OLAP [4] performance of PostgreSQL and MongoDB on six distributed configurations deployed on OpenStack [5], by varying the database size and the number of data distribution nodes.

For this purpose, we used the TPC-H Benchmark [6] database and tools, to populate the TPC-H database tables with a specified volume of data. We started with the original TPC-H schema, devising a module to populate a corresponding MongoDB schema by transforming (mainly by nesting) the data structure. This transformation allows a

MongoDB collection to integrate data originally scattered across two or more tables.

We developed a set of 296 SQL queries (for the original TPC-H schema) that vary in terms of number of joins, filter clauses, or group clauses. Subsequently, this query set was translated into MongoDB's Aggregation Framework (MAF) and adapted to the new document database structure. In this new structure, some of the joins between collections become unnecessary, as the data is denormalized and stored as nested arrays inside documents.

The queries were run on both PostgreSQL and MongoDB setups across six scenarios. These scenarios were defined by the combination of two small-scale factors (of 0.01GB and 0.1GB) and three distribution architectures with 3, 6, and 9 nodes. A 10-

minute timeout was imposed for the execution of each query. Queries that did not complete within 600 seconds were canceled.

The preliminary results on the relationship between query completion (for both database servers) and the database size and the distribution setup were presented in [6]. In this paper, we examine the relationships between the query completion and some parameters describing the query complexity; we also developed, tuned, and interpreted Machine Learning (ML) classification models for predicting query completion on various predictors about database size, distribution setup and query complexity. The remaining of this paper is organized as follows: section 2 examines the results of some previous studies approaching the performance comparison of SQL (PostgreSQL) and document (MongoDB) data stores. Section 3 provides a brief overview of the distributed architectures of Citus (a distributed PostgreSQL solution) and MongoDB. Section 4 describes the experimental setup, data analysis methods and tools used for obtaining the results. Results are analyzed in section 5. The paper concludes with the main findings, pointing out the study's main limitations, and some future directions for research.

## 2 Previous studies on OLAP performance comparison of SQL and NoSQL database servers

Both PostgreSQL and MongoDB are highly popular on the database market [7] and a considerable body of literature targeted their OLTP (On-Line Transaction Processing) and/or OLAP performance comparison. Results are far from convergent, as seen below.

In [8] PostgreSQL outperformed MongoDB. Güney and Ceylan [9] found that MongoDB had certain advantages in specific cases, particularly in data sorting operations, while PostgreSQL performed

better in bulk data extraction and more complex query operations. Villalobos et al. [10] found that PostgreSQL performs more efficiently with complex queries involving intersection or combination functions, while MongoDB is better suited for simpler queries involving only filtering of descriptive, non-geographic data. In [11] MongoDB excelled in managing national election data, with consistent execution times regardless of operation size. PostgreSQL also performed well, with memory usage increasing linearly with the size of the operation. Also, in [12] MongoDB significantly outperformed SQL Server in processing e-government data.

Yedilkhan et al. [13] showed that both MongoDB and PostgreSQL are suitable for particular use cases and scenarios. The comparison was based on performance evaluations conducted across multiple clusters, including both cloud and on-premises environments. In [14] MongoDB outperformed PostgreSQL in terms of latency. Makris et al. [15] concluded that PostgreSQL performed better than MongoDB when managing spatio-temporal data. In [16] results showed that relational databases are more efficient than non-relational databases when executing the basic types of database operations (CRUD – Create, Read, Update, Delete).

Tracz and Plechawska-Wójcik [17] analyzed the performance of relational and non-relational databases using MS SQL Server, MongoDB, and CouchDB. A total of 30 test series were conducted for each database server across the scenarios of interest, using the following record counts: 500, 1000, 2000, 5000, and 10000; results revealed that SQL Server ranks first, followed by MongoDB in the second place, and CouchDB as the least efficient of the three. In [18] results indicated that NoSQL databases are a better alternative to their relational counterparts. Setyawan et al. [19] found that MongoDB is well-suited for handling large volumes of data, while MySQL recorded optimal performance for queries executed on a smaller scale.

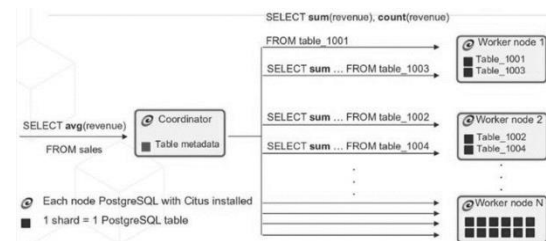
Figueiredo et al. [20] showed that PostgreSQL manages better data that is well structured, due to its robust support for complex queries. In their study, a 10 GB dataset was processed by each database server, with three types of queries: 1st, 5th and 6th (from the TPC-H benchmark query-set) selected for performance evaluation. Two testing scenarios were employed, one involving approximately 750 MB of data inserted into PostgreSQL and MongoDB, and another with a 10 MB file inserted into HarperDB. On the response time, PostgreSQL recorded the best database performance.

Sals et al. (2023) [21] argued that MongoDB outperforms MySQL in data storage and processing capabilities. Moreover, Antas et al. [22] concluded that MongoDB has better results for almost all large data volume tests; similarly, in [23] MongoDB performed better for CRUD operations. Abukabar et al. [24] designed an evaluation benchmark and identified that MongoDB recorded superior performance in handling CRUD operations, while MySQL excels in executing stored procedures. Matallah et al. [25] found that MongoDB recorded smaller execution time than MySQL for all types of query operations. In [26] the results showed that MongoDB had significantly higher throughput than both MySQL and PostgreSQL, particularly for small transaction sizes, indicating that MongoDB is a better choice for applications which involve heavy writing of indexed spatial data.

Reichardt et al. [27] compared the performance of NoSQL data stores, relative to their relational/SQL counterparts. They found that MongoDB is more suitable than MySQL for a wide range of dataset sizes. Naufal et al. [28] found MongoDB outperforms in the Update and Delete operations based on runtime differences.

### 3 Distributed architectures for PostgreSQL/Citus and MongoDB. Deployment on OpenStack

Citus is an open-source and easy-to-deploy PostgreSQL extension which enhances PostgreSQL by distributing the database using horizontal scaling and parallel query processing [29]. It utilizes multiple PostgreSQL instances to distribute both data and queries across the nodes (in this paper we deployed distributed setups with 3, 6 and 9 nodes). Thus, it provides efficient management of large datasets and high query volumes. In a Citus cluster (see Fig. 1), data is divided into shards and distributed across worker nodes. The master node handles the metadata for the distributed tables, coordinates query execution by delegating tasks to worker nodes, and aggregates the results. Worker nodes store the data shards and execute the queries [30].



**Fig. 1.** Citus architecture ([29])

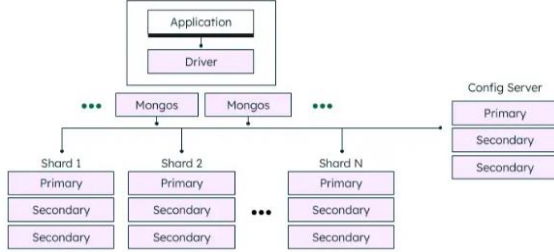
This architecture enables Citus to uphold the robustness and flexibility of PostgreSQL. At the same time, it delivers substantial performance improvements for OLAP workloads [31].

MongoDB is the most popular NoSQL product [7] which proved to be a reliable solution in providing database scalability, flexibility, and high performance. The database uses JSON-like documents with flexible schemas, which makes it easier to manage various data types and speeds up the development process.

Moreover, it employs a sharding architecture (Fig.2) to spread the processing load and storage across multiple servers. By horizontal scalability, MongoDB can handle large volumes of data with good overall performance in data processing. Data is split



into smaller, more manageable units known as shards, which can be hosted on separate servers or containers for better and easier management.



**Fig. 2.** MongoDB architecture ([32])

MongoDB distributes queries across these shards, allowing it to efficiently handle large data volumes and high traffic loads [32]. The data chunks are duplicated among shards to ensure high availability. Mongos instances route queries and manage data distribution, while config servers handle metadata and configuration settings.

MongoDB supports replication through replica sets, which further enhances data availability and fault tolerance in critical systems. Each replica set includes a primary node for writes and multiple secondary nodes that maintain copies of the data, these are in pairs of 3 machines in most cases. If failure occurs, one of the secondary nodes is automatically promoted to primary, ensuring minimal downtime and data consistency.

More details on Citus and MongoDB distributed setups are provided in section 4.3.

#### 4 The experimental design. Research method and tools.

We created an initial set of 296 SQL queries (for the TPC-H benchmark database) in PostgreSQL and translate them into MongoDB Aggregation Framework queries on a JSON schema with nested arrays in collections (as described in section 4.1). The queries were executed multiple times on both servers, with a 600-second timeout, varying the database size (scale factor)

and the data distribution (the number of nodes).

Whereas most of the previous studies presented in section 2 assessed the database performance with the query response times metric, our research focuses on the successful completion of queries within a timeout period of 10 minutes. In [6] the successful query completion was examined in relation to the database server and the nodes of data distribution using inferential statistics. In this paper we approached a series of research questions and sub-questions, also concerning the query successful completion, as follows:

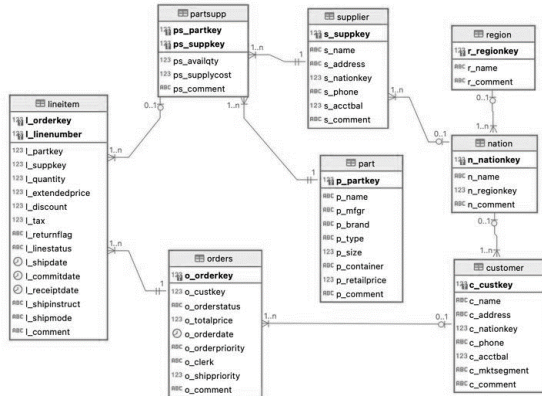
- RQ1: Is the query completion/success associated with the parameters describing the query complexity?
  - RQ1a: Is query completion associated with the number of joins/lookups in the query?
  - RQ1b – Is query completion associated with the number of filters (predicates for record selection)?
  - RQ1c – Is query completion associated with the record grouping?
  - RQ1d – Is query completion associated with limiting the number of records in the result?
- RQ2: Can the query successful execution be predicted based on database size, physical setup and query complexity parameters?
  - RQ2a: Do Machine Learning classification models reliably predict the success of query execution based on the predictors related to the data volume, data distribution and query complexity?
  - RQ2b: Which are the most important variables for the prediction of query completion? How their distribution influences the query success?

##### 4.1 Database structure in PostgreSQL and MongoDB

Since our study is focused on the OLAP performance [33] of MongoDB and PostgreSQL, we relied on the TPC-H benchmark database. This benchmark evaluates database systems performance by

using a series of business-oriented queries (which we did not use in this paper).

In PostgreSQL the database followed the original TPC-H schema consisting of eight tables, as represented in Fig. 3. For details on the table attributes and constraints, see [34] and [6].



**Fig. 3.** TPC-H database schema ([34] [6])

When devising the corresponding MongoDB collections, we did not simply map the PostgreSQL tables into “flat” collections. Instead, we combined some tables through nesting and thus we incorporate some redundancy into the database. The resulting collections were:

- *customer\_nation*
- *lineitem*
- *lineitem\_orders*
- *lineitem\_partsupp*
- *part*
- *partsupp*
- *region*
- *supplier\_nation*.

Collections with simple names mirror the PostgreSQL tables directly, containing a straightforward JSON representation with the same attributes. The compounded collections, identified by names with underscores, merge data from two PostgreSQL tables. These collections are designed to reflect more complex relationships by nesting documents and arrays. As illustrated in Fig. 4, each document in the *lineitem\_orders* collection represents an order from the *orders* table which includes (lines 7 to 29

in Fig. 4) an array of line items associated with that order, integrating data from both the *orders* and *lineitem* tables.

```

1 { // this is the first document (describing the first order)
2   "o_orderkey": 1, "o_custkey": 370, "o_orderstatus": "O",
3   "o_totalprice": 172799.49,
4   "o_orderdate": {"$date": "1996-01-02T00:00:00.000-0700"},
5   "o_orderpriority": "5-LOW", "o_clerk": "Clerk#00000951",
6   "o_shippriority": 0, "o_comment": "nstructions sleep furiously among ",
7   "lineitem": [
8     { // this is the first line of the current order
9       "l_orderkey": 1, "l_partkey": 1552, "l_suppkey": 93,
10      "l_linenum": 1, "l_quantity": 17.00, "l_extendedprice": 24710.35,
11      "l_discount": 0.04, "l_tax": 0.02, "l_returnflag": "N",
12      "l_linestatus": "O", "l_shipdate": {"$date": "1996-03-13T00:00:00.000-0700"},
13      "l_commitdate": {"$date": "1996-02-12T00:00:00.000-0700"},
14      "l_receiptdate": {"$date": "1996-03-22T00:00:00.000-0700"},
15      "l_shipinstruct": "DELIVER IN PERSON",
16      "l_shipmode": "TRUCK", "l_comment": "regular courts above the"
17    }
18  ],
19  { // this is the second line of the current order
20    "l_orderkey": 1, "l_partkey": 674, "l_suppkey": 75,
21    "l_linenum": 2, "l_quantity": 36.00, "l_extendedprice": 56688.12,
22    "l_discount": 0.09, "l_tax": 0.06, "l_returnflag": "N",
23    "l_linestatus": "O", "l_shipdate": {"$date": "1996-04-12T00:00:00.000-0700"},
24    "l_commitdate": {"$date": "1996-02-28T00:00:00.000-0700"},
25    "l_receiptdate": {"$date": "1996-04-28T00:00:00.000-0700"},
26    "l_shipinstruct": "TAKE BACK RETURN",
27    "l_shipmode": "MAIL", "l_comment": "ly final dependencies: slyly bold "
28  }
29  ] // ... the remaining lines of the current order
30 }
31 //, ... the remaining orders

```

**Fig. 4.** Structure of a document in collection *lineitem\_orders* ([6])

In PostgreSQL the join is required when extracting data from both *lineitem* and *orders* tables; but in MongoDB's Aggregation Framework, this join (*\$lookup* function) is not needed because all the relevant data is contained within a single collection.

In MongoDB the names of the collections which map two PostgreSQL tables are formatted with the child table's name on the left side of the underscore and the parent table's name on the right side. In collection *lineitem\_orders*, *lineitem* is the child table and table *orders* is the parent. Each document in this collection corresponds to a record from the parent table (*orders*), and within each document, there is an array that stored all related records from the child table (*lineitem*).

## 4.2 SQL and Aggregation Framework Queries

The initial 296-query set (written in the PostgreSQL dialect of SQL) was devised to ensure some variability on the parameters describing the query complexity, as contained in the SELECT, FROM, WHERE, GROUP BY, ORDER BY, LIMIT/FETCH and OFFSET/SKIP clauses. The query set

was converted into the syntax of the main declarative query language in MongoDB – the Aggregation Framework, considering the new (nested) JSON structure.

Table 1 illustrates two scenarios of mapping SQL queries to the Aggregation Framework language, considering the MongoDB's “nested” structure (using arrays). In SQL, the 187th query involves

a join between *customer*, *orders*, *lineitem* and *partsupp* tables. For the same query in MongoDB, Aggregation Framework uses the *\$lookup* method to achieve a similar effect by joining data from different collections. By contrast, the 277th query requires join only in SQL, because in Aggregation Framework there is no need for join since all query data are stored in a single collection, *customer\_nation*.

**Table 1.** Two examples of query mapping

Query/ Syntax	Query content
Q187 in SQL	<pre>select customer.c_name, orders.o_orderdate, lineitem.l_quantity,partsupp.ps_availqty from customer join orders on customer.c_custkey = orders.o_custkey join lineitem on lineitem.l_orderkey = orders.o_orderkey join partsupp on lineitem.l_partkey = partsupp.ps_partkey where l_quantity &lt; 5 and l_tax &gt; 0.05 and l_discount &gt; 0.08 limit 800;</pre>
Q187 in AF	<pre>db.customer_nation.aggregate([{\$unwind: "\$customer"}, {\$lookup: {from: "lineitem_orders", localField: "customer.c_custkey", foreignField: "o_custkey", as: "orders_1join"}}, {\$unwind: "\$orders_1join"}, {\$lookup: {from: "lineitem",localField: "orders_1join.o_orderkey", foreignField: "l_orderkey", as: "lineitem_2join"}}, {\$match: {"lineitem_2join.l_quantity": {\$lt: 5}, "lineitem_2join.l_tax": {\$gt: 0.05}, "lineitem_2join.l_discount": {\$gt: 0.08}}}, {\$lookup: {from: "partsupp",localField: "lineitem_2join.ps_partkey", foreignField: "l_partkey",as: "partsupp_3join"}}, {\$project: { _id: 0, c_name: "\$customer.c_name", o_orderdate: "\$orders_1join.o_orderdate", l_quantity: "\$lineitem_2join.l_quantity", ps_availqty: "\$partsupp_3join.ps_availqty"}}, {\$limit: 800}]);</pre>
Q277 in SQL	<pre>select count(c_custkey), n_name from customer join nation on customer.c_nationkey = nation.n_nationkey where c_acctbal &gt; 0 and c_mktsegment = 'FURNITURE' and n_regionkey = 1 group by n_name;</pre>
Q277 in AF	<pre>db.customer_nation.aggregate([ {\$unwind: "\$customer"}, {\$match: {"customer.c_acctbal": { \$gt: 0 }}}, {\$match: {"customer.c_mktsegment": "FURNITURE "}}, {\$match: {"n_regionkey": 1}},{\$group: {_id: "\$n_name",NumberOfCustomer: { \$sum: 1 }}}, {\$project: {_id: 0, n_name: "\$_id",NumberOfCustomer:1}} ]);</pre>

When translating the SQL queries into MongoDB Aggregation Framework queries for yielding the equivalent results, we strived to maintain simplicity in the syntax for both languages.

### 4.3 Physical Setup

The experimental setup was deployed on the RaaS-IS platform, which operates as a private cloud managed by OpenStack, as described in [6]. The RaaS-IS datacentre

architecture includes 20 servers: 3 controller nodes to ensure high availability of OpenStack services, 16 compute nodes for Virtual Machine and container provisioning, and a management server responsible for MAAS, Juju, LDAP, and network management. The storage infrastructure uses an HPE 3PAR 8440 SAN with a total raw capacity of 760 TB, of which 550 TB is usable, and includes an 8 TB SSD cache for enhanced performance. The SAN's NL-SAS HDDs are set up in RAID 5, while the SSD cache uses RAID 6 to deliver higher performance. The compute nodes have dual Intel Xeon Gold 6240 CPUs (18 cores, 2.6 GHz each), 128 GB of RAM@2933MHz, and 300 GB SAS drives in RAID 1. Together, these provide 1152 virtual cores and 2 TB RAM for the RaaS-IS project and infrastructure.

#### 4.3.1 PostgreSQL Citus Cluster Setup

The Citus cluster was configured and partially deployed using the OpenStack Heat orchestration module, which allows for Infrastructure as Code (IaC) using built-in OpenStack tools. The cluster was set up on virtual machines within the cluster nodes running a custom-configured Ubuntu Linux 18.04.6 LTS (kernel version 4.15.0-213-generic), along with TPC-H v3.0 and Citus v11.3. To streamline cluster management, custom bash scripts were developed to enhance monitoring, result retrieval, and automated deployment.

The Citus cluster consisted of 10 machines, designed to provide distributed parallel query execution and scalability for analytical workloads. The cluster setup was as follows:

- Coordinator Node (1 machine): This node served as the primary entry point for queries, distributing them across worker nodes while maintaining metadata and query routing.
- Worker Nodes (9 machines): These nodes handled data storage and

parallel execution of queries, improving performance for analytical workloads

#### 4.3.2 MongoDB Sharded Cluster setup

The MongoDB cluster was configured and partially configured using the OpenStack Heat orchestration module. This module uses built-in OpenStack tools to allow Infrastructure as Code (IaC). The MongoDB cluster was set up on virtual machines within the compute nodes, using a custom-configured Ubuntu Linux 18.04.6 LTS (kernel version 4.15.0-213-generic), TPC-H v3.0, and MongoDB v6.0.14. A set of bash scripts was developed to improve monitoring and configuration for retrieving results and automating the cluster setup.

The MongoDB cluster comprised 13 machines configured to ensure optimal performance and reliability, mirroring a typical production environment.

The servers were organized as follows:

- Router Node (1 machine): This server directed client requests to the correct shard.
- Config Nodes (3 machines) – these nodes managed the cluster's metadata and configuration, ensuring smooth operation.
- Shards (9 machines): The data was spread across three shards, each with three replica sets. Each set included a primary node for write operations and two secondary nodes for redundancy and high availability. The number of shards could be adjusted depending on the test scenario.

Each machine was set up with 4 cores, 4GB of RAM, and a 200GB storage volume on the SAN.

#### 4.4 Variables (outcomes and predictors)

Table 2 shows the variables used in the analysis. The first three variables define the physical setup of each scenario in the experiment. Variable *success* refers to the outcome of the query execution (whether it was completed within the 10-minute timeout).

**Table 2.** Variable description

Variable	Description
<i>db_server</i>	name of the DBMS
<i>scale_factor</i>	size of the DBMS
<i>n_of_nodes</i>	number of nodes used for data distribution
<i>success</i>	binary variable whose value is TRUE when the query was successfully completed within the 10-minute timeout, and FALSE otherwise
<i>join</i>	number of JOIN clauses (PostgreSQL), equivalent to \$lookup (MongoDB)
<i>where</i>	number of WHERE clauses (PostgreSQL), equivalent to \$match (MongoDB)
<i>count</i>	number of COUNT clauses (PostgreSQL), equivalent to count: (MongoDB)
<i>group_by</i>	number of GROUP BY clauses (PostgreSQL), equivalent to \$group (MongoDB)
<i>having</i>	number of HAVING clauses (PostgreSQL), equivalent to \$match (MongoDB)
<i>limit</i>	number of LIMIT clauses (PostgreSQL), equivalent to \$limit (MongoDB)

The variable names related to the query complexity are inspired from the SQL syntax, and the second column points to the Aggregation Framework equivalent feature.

#### 4.5 Method and tools for data analysis

For RQ1a-RQ1c the analysis relied on classical inferential statistics. For RQ2a and RQ2b we built, tuned and interpreted a series of ML classification models based on two of the most popular classification algorithms, Random Forest and Extreme Gradient Boosting (XGBoost) algorithms.

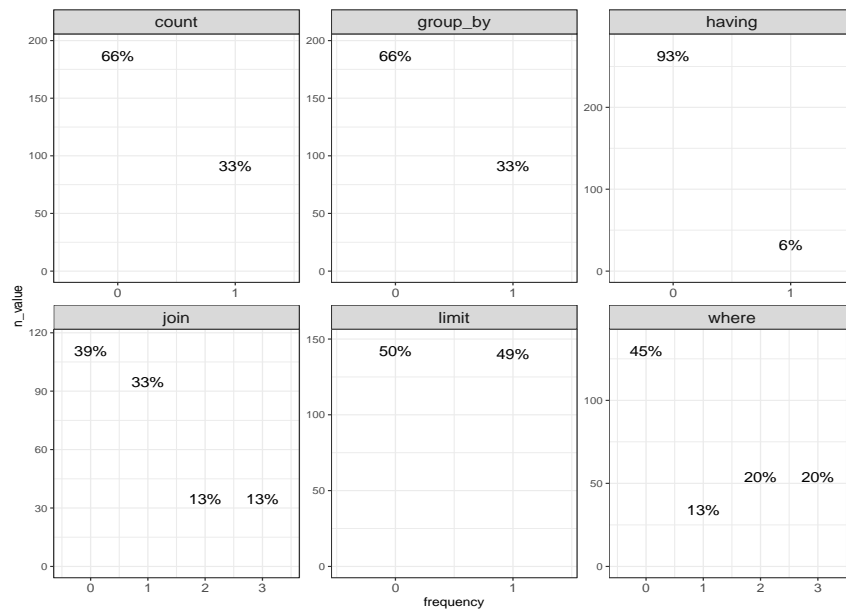
After the query results and parameters were collected and integrated, the dataset was examined using Exploratory Data Analysis. The Chi-square test of independence was applied to assess the association between nominal variables. We employed Random Forest and XGBoost to identify the most important predictors for the success of query execution.

The dataset was processed and analyzed using R programming language [35], employing tools from the *tidyverse* package collection [36] along with *ggstatsplot* [37]. The *tidyverse* suite is a set of R packages designed for data manipulation and visualization, while *ggstatsplot* and *rstatix* provide additional functionalities for statistical analysis and visualization of results. All models were built and tuned with the *tidymodels* ecosystem [38].

All the Interpretable ML techniques (Variable Importance, Partial Dependency Profiles, Individual Conditional Expectation, and Accumulated Local Effects Profiles) were deployed using the *DALEX* ecosystem ([39]).

## 5 Results and discussion

All 296 queries were run on both database servers across six different scenarios. The chart in Fig. 5 illustrates the distribution of query parameters, providing a glimpse into the queries complexity.



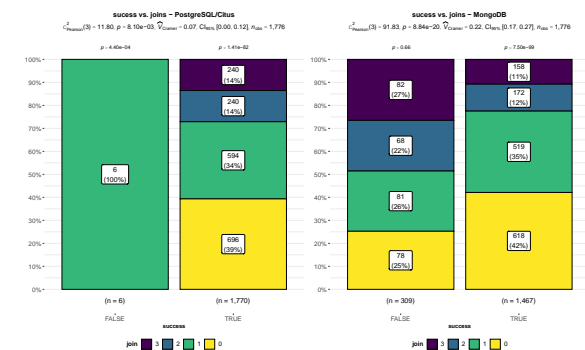
**Fig. 5.** Distribution of query parameters

The chart reveals that 33% of the queries involved grouping records and performing aggregation by counting the records within each group, 45% had no filters applied, and roughly half of the queries imposed a limit on the result size. As presented in [6], there is a correlation between query completion (within the 10-minute timeout) and the database server. Of the 3552 queries executed across both servers in all six scenarios, 315 failed to complete. Of these, 309 failed on MongoDB, while only six failed on PostgreSQL/Citus. Of the 3237 successful queries, 1770 (55%) were executed on Citus, with the remaining 1467 (45%) completed on MongoDB.

Also, [6] found that the relationship between query completion (within the 10-minute timeout) and the data distribution setup differs between database servers. In Citus, the proportion of successful queries remains consistent across the three data distribution scenarios. However, MongoDB shows more variability: the 6-node setup had the highest failure rate, while the 9-node setup performed the best.

The subsequent series of statistical tests concerned the associations between the successful query completion and some variables describing the query

complexity. To start with *RQ1a*: (*Is query completion associated with the number of joins/lookups in the query?*), Fig. 6 shows the results of the Chi-Square test of independence between variables *success* and *joins* for PostgreSQL/Citus (left) and MongoDB (right).



**Fig. 6.** Association query success vs. number of joins

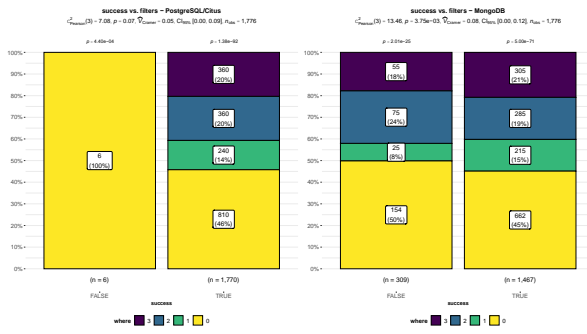
In Citus, successful and unsuccessful queries recorded a similar average number of joins (1). By contrast, in MongoDB, the joins/lookups appear more costly since for the completed queries (1467), the average number of joins was 0.91, whereas, for the canceled queries (309), the average number of joins was 1.50.

For Citus, the Chi-square test of independence indicated a p-value below 0.05 and an effect size (*Cramer's V*) of 0.07,



providing evidence to state that the number of joins is significantly linked to the successful completion of queries within the 10-minute timeout. Meanwhile, for MongoDB, the Chi-square test computed a p-value less than 0.05 and a *Cramer's V* effect size of 0.22, which offer support to state that the number of joins is significantly related with the successful completion of queries within the 10-minute timeout.

Regarding *RQ1b* (*Is query completion associated with the number of filters/predicates for record selection?*), Fig.7 shows the results of the Chi-Square test of independence between variables *success* and *filters* for Citus (left) and MongoDB (right).



**Fig. 7.** Association success vs. number of filters

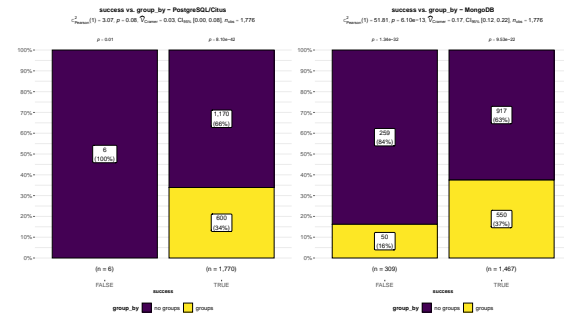
In Citus, for the successful queries (1770), the average number of filters was 1.14 and for the unsuccessful queries (6) the average number of filters was 0. In MongoDB, successful and unsuccessful queries recorded a similar average number of filters, ~ 1.

For Citus, the Chi-square test recorded a p-value greater than 0.05 and an effect size (*Cramer's V*) of 0.05, suggesting that the number of filters is not associated with the successful completion of queries. Similarly, for MongoDB, the Chi-square computed a p-value smaller than 0.05 and an effect size (*Cramer's V*) of 0.08, indicating that the number of predicates is not significantly associated with the successful completion of queries, just as

observed in Citus.

Addressing the next research question, *RQ1c* - *Is query completion associated with grouping (aggregation) the records?*, Fig. 8 shows the findings of the Chi-Square test of independence between the variables *success* and *group\_by* for both Citus (left) and MongoDB (right).

In Citus, the unsuccessful queries (6) included those that were executed without any record grouping. In the case of successful queries (1770), 66% were executed without record grouping, whereas 34% included it. In contrast, MongoDB experienced 309 unsuccessful queries, the majority of which did not feature record grouping, while out of 1467 successful queries, only 37% included record grouping.



**Fig. 8.** Association success vs. record grouping

For Citus, the Chi-square test computed a p-value greater than 0.05 and an effect size (*Cramer's V*) of 0.03, which fail to support a significant association between record grouping and query completion. Conversely, for MongoDB, the Chi-square test of independence showed a p-value less than 0.05 and a higher effect size (0.17), which points to a strong association between record grouping and query completion.

Regarding the subsequent research question, *RQ1d* - *Is query completion associated with fixing a limit of records in the result?*, Fig. 9 presents the outcomes of the Chi-Square test of independence between variables *success* and *limit* for Citus (left) and MongoDB (right).

In Citus, the unsuccessful queries (6) were

those executed without limiting the result size, whereas for the successful queries (1770), more than a half of queries were executed without result size limitation. In contrast, MongoDB recorded 309 unsuccessful queries, the majority of which did not impose a limit on the result size. Among the 1467 successful queries, less than a half of the total applied a size limitation.



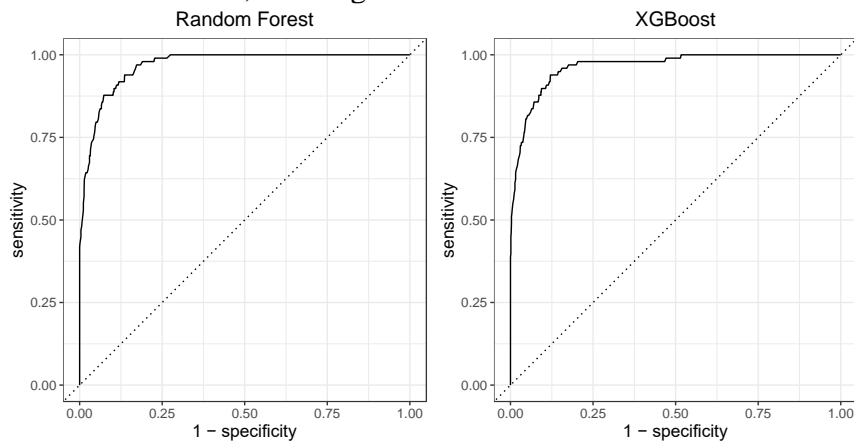
**Fig. 9.** Association of success and limiting the result size

For Citus, the Chi-square test of independence revealed a p-value greater than 0.05 and an effect size (*Cramer's V*) of 0.03, so limiting the result size seems not to be associated with the successful completion of queries. Alternatively, for MongoDB, the p-value of the test was below 0.05 and the effect size (0.17) was higher; for this database server, limiting

the result size seems associated with the query completion.

Next, we examined whether the predictors in Table 2 can reliably predict the query completion. We built and tuned a series of ML classification models, with *query\_completion* (variable that recorded the query success) as the outcome (target) and the remaining variables in Table 2 as the predictors. As described in Section 4.5, only two classification algorithms were employed for this paper, Random Forest (RF) and XGBoost. The metric performance for selecting the best model was *ROC\_AUC*. Best RF performance was recorded for the hyper-parameter combination of *mtry* = 9 and *min\_n* = 13, whereas for the best hyper-parameter combination for XGB was *mtry*=7, *min\_n*=5, *tree\_depth*=9, *learn\_rate*=0.0618, *loss\_reduction*=0.0000000189, and *sample\_size* =0.509.

Fig.10 displays the performance of both algorithms on the new data (the test set). RF slightly overperformed XGB, with a *ROC\_AUC* of 0.968 (relative to 0.965 for XGB). Recorded accuracy was higher for XGB - 0.950 (relative to 0.945 for RF). Consequently, for further estimation of predictor's importance, the RF best model was preferred.



**Fig. 10.** Performance of the tuned classification models on the test set

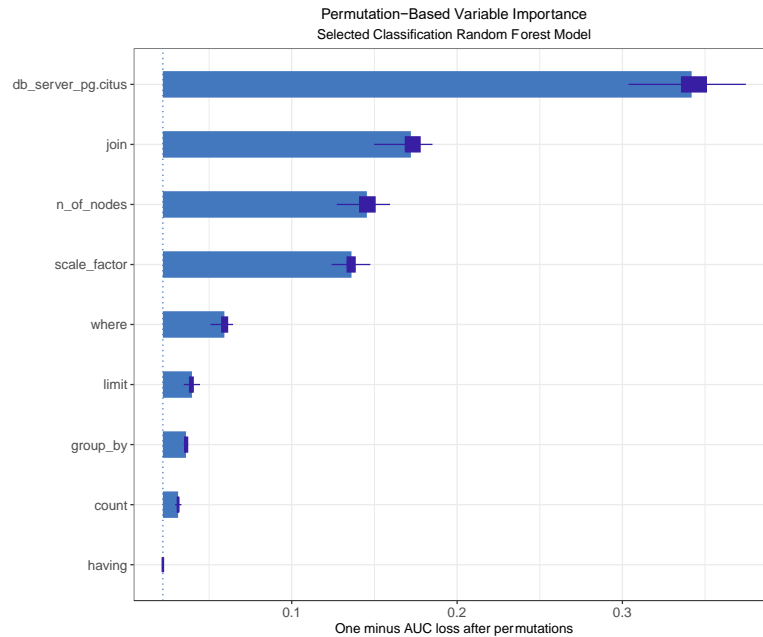
Fig. 11 displays the permutation-based importance of variables as estimated by the RF algorithm. The graphical representation shows that the variable

*db\_server\_pg.citus* has the highest importance, as permuting this variable results in the most significant performance loss, followed by the variables *join*,



*scale\_factor* and *n\_of\_nodes*. Variables *group\_by*, *having*, and *count* have the

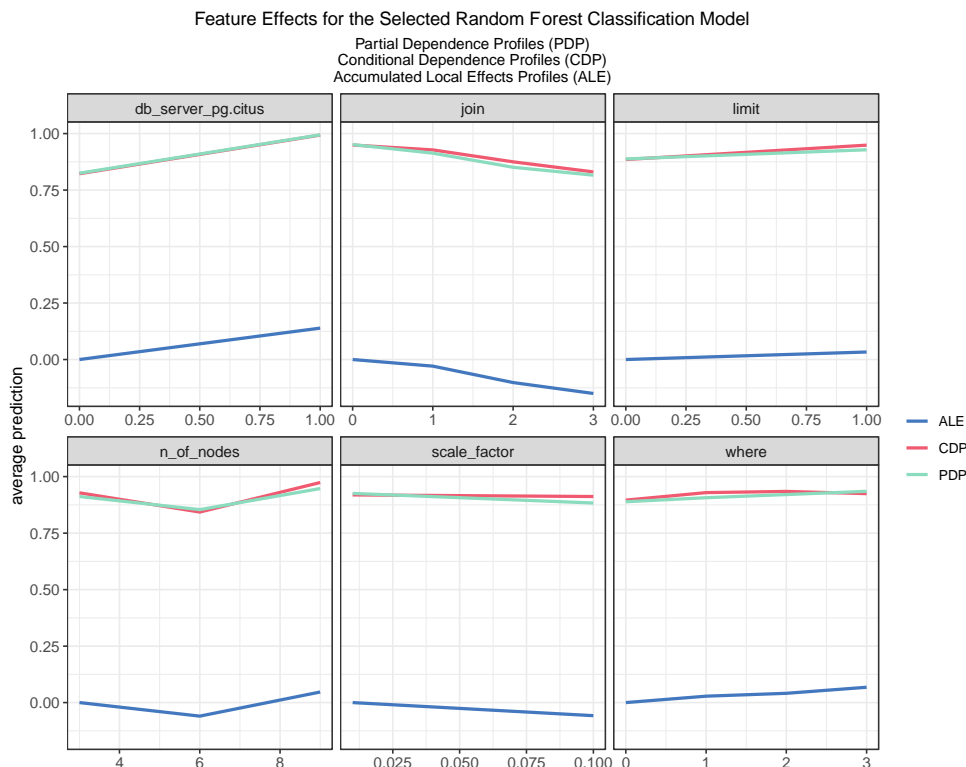
lowest importance with a minimal impact on the model's performance.



**Fig. 11.** Permutation-based variable importance on RF classification model

The six most influential variables in Fig.11 were selected for computing and displaying their effects on the average prediction of the outcome probability (probability of query to be successfully

completed), using some techniques of interpretable ML. Fig. 12 shows three metrics describing the predictor's effect, Partial Dependence Profiles (PDP), Conditional Dependence Profiles (CDP) and Accumulated Local Effects Profiles (ALE).



**Fig. 12.** Model explanations for the selected RF classification model.

Regarding the *db\_server\_pg.citus* variable, the average prediction of query completion probability appears to be positively associated with the use of the Citus database server, as indicated by the upward trend of the curve. For the *join* variable, the average prediction of query completion probability decreases as more joins are added in constructing the query. Regarding variable *limit*, setting a limited size of records contributes to achieving a higher average prediction. As for the *n\_of\_nodes* variable, employing either 3 or 9 nodes increase the odds of success for query completion. Unexpectedly, the influence of the database size (*scale\_factor*) does not seem to differ between the two levels relatively, which can be explained by the small scales of 0.01 and 0.1 GB used. As for variable *where*, the average prediction of query completion increases when additional where clauses are added into the query.

## 6 Conclusion

Consistent with some of the earlier research, our findings indicate that distributed PostgreSQL/Citus outperforms MongoDB regarding query performance for smaller databases, even when both systems are subjects of the same requirements and physical resources. Also, we identified that the most critical factor influencing query execution is the type of database server, followed by the complexity of joins used in query construction and the data distribution setup.

The primary limitation of the experimental setup is the small database size, which, at 0.01 GB and 0.1 GB, is modest by OLAP standards. Additional limitations include the relatively small query set (296 queries), the complexity of the queries, and the brief 10-minute timeout imposed for query execution.

Future experiments on the setup should involve processing larger datasets and employing different data nesting

structures in MongoDB collections. Moreover, increasing the query set size, expanding the variability of query parameters, and exploring various data distribution scenarios could provide more comprehensive insights.

## 7 Acknowledgment

Data processing and analysis for this study were supported by Competitiveness Operational Program Romania under project SMIS 124759 - RaaS-IS (Research-as-a-Service Iasi).

## References

- [1] C. Madera and A. Laurent, "The next information architecture evolution: the data lake wave," in *Proc. 8th Int. Conf. on Management of Digital EcoSystems (MEDES)*, 2016, pp. 174-180. doi: 10.1145/3012071.3012077.
- [2] F. Naregsian, E. Zhu, R. Miller, K. Pu, and P. Arocena, "Data Lake Management: Challenges and Opportunities," *Proc. VLDB Endow.*, vol. 12, no. 12, pp. 1986-1989, 2019. doi: 10.14778/3352063.3352116.
- [3] P. Pedreira, O. Erling, K. Karanasos, S. Schneider, and W. McKinney, "The Composable Data Management System Manifesto," *Proc. VLDB Endow.*, vol. 16, no. 10, pp. 2679-2685, 2023. doi: 10.14778/3603581.3603604.
- [4] D. Martinez-Mosquera, "Integrating OLAP with NoSQL Databases in Big Data Environments: Systematic Mapping," *Big Data Cogn. Comput.*, vol. 8, no. 6, 2024.
- [5] K. Prabhakar, J. Kurunandan, A. Amjad, P. Prabu, and B. Rajkumar, "OpenStackDP: A scalable network security framework for SDN-based OpenStack cloud infrastructure," *J. Cloud Comput.: Adv. Syst. Appl.*, vol. 12, no. 1, 2023. doi: 10.1186/s13677-023-00406-w.
- [6] M. Fotache, C. Badea, M. I. Cluci, C. Pinzaru, C. S. Esanu, and O. Rusu, "OLAP Performance of Distributed PostgreSQL and MongoDB on

- OpenStack. Preliminary Results on Smaller Scale Factors,” in *Proc. 23rd RoEduNet Conf.: Networking in Education and Research*, 2024, doi: 10.1109/RoEduNet64292.2024.10722556
- [7] DB Engines Ranking [Online]. <https://db-engines.com/en/ranking>.
- [8] A. Makris, K. Tserpes, G. Siliopoulos, D. Zissis, and D. Anagnostopoulos, “MongoDB vs PostgreSQL: A comparative study on performance aspects,” *Geoinformatica*, vol. 25, pp. 243–268, 2021. doi: 10.1007/s10707-020-00407-w.
- [9] E. Güney and N. Ceylan, “Response Times Comparison of MongoDB and PostgreSQL Databases in Specific Test Scenarios,” in *ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, 2022, pp. 178–188.
- [10] M. T. Villalobos, L. V. Acuna, and R. Q. Oviedo, “Comparison of the Response Times of MongoDB and PostgreSQL According to Type of Query in Geographical Databases,” *Computación y Sistemas*, vol. 24, no. 4, pp. 1461–1469, 2020.
- [11] L. G. Wiseso, M. Imrona, and A. Alamsyah, “Performance Analysis of Neo4j, MongoDB, and PostgreSQL on 2019 National Election Big Data Management Database,” in *International Conference on Science in Information Technology*, Palu, Indonesia, 2020.
- [12] A. Flores, S. Ramírez, R. Toasa, J. Vargas, R. U. Barrionuevo, and J. M. Lavin, “Performance Evaluation of NoSQL and SQL Queries in Response Time for the E-government,” in *2018 International Conference on eDemocracy & eGovernment (ICEDEG)*, 2018, pp. 257–262. doi: 10.1109/ICEDEG.2018.8372362.
- [13] Y. D. Mukasheva, A. Bissengaliyeva, D. Suynullayev, “Performance Analysis of Scaling NoSQL vs SQL: A Comparative Study of MongoDB, Cassandra, and PostgreSQL,” in *2023 IEEE International Conference on Smart Information Systems and Technologies (SIST)*, Astana, Kazakhstan, 2023, pp. 479–483. doi: 10.1109/SIST58284.2023.10223568.
- [14] M. M. Eyada, W. Saber, M. M. El Genidy, and F. Amer, “Performance Evaluation of IoT Data Management Using MongoDB Versus MySQL Databases in Different Cloud Environments,” *IEEE Access*, vol. 8, pp. 110656–110668, 2020. doi: 10.1109/ACCESS.2020.3002164.
- [16] W. Ali, M. U. Shafique, M. A. Majeed, and A. Raza, “Comparison between SQL and NoSQL databases and their relationship with Big Data analytics,” *Asian Journal of Research in Computer Science*, vol. 4, no. 2, pp. 1–10, 2019. Article no. AJRCOS.51946, ISSN: 2581-8260.
- [15] A. Makris, K. Tserpes, G. Spiliopoulos, and D. Anagnostopoulos, “Performance evaluation of MongoDB and PostgreSQL for spatio-temporal data,” in *EDBT: 22nd International Conference on Extending Database Technology*, Lisbon, Portugal, 2019, pp. 1–9.
- [16] R. Čerešňák and M. Kvet, “Comparison of query performance in relational and non-relational databases,” in *Proc. of the 13th Scientific Conference on Sustainable, Modern and Safe Transport (TRANSCOM 2019)*, Nový Smokovec, vol. 40, pp. 170–177, May 29–31, 2019. doi: 10.1016/j.trpro.2019.07.027.
- [17] P. M. Tracz and M. Plechawska-Wójcik, “Comparative analysis of the performance of selected database management system,” *Journal of Computer Sciences Institute*, vol. 31, pp. 89–96, 2024. doi: 10.35784/jcsi.5927.
- [18] A. M. Kausar and M. Nasar, “SQL Versus NoSQL Databases to Assess Their Appropriateness for Big Data

- Application,” *Recent Advances in Computer Science and Communications*, vol. 14, no. 4, 2021, doi: 10.2174/2213275912666191028111632
- [19] A. B. Setyawan, I. A. Kautsar, and N. L. Azizah, “Query Response Time Comparison SQL and No SQL for Contact Tracing Application,” in *Proceedings of the 4th Seminar Nasional Sains. Procedia of Engineering and Life Science*, vol. 2, no. 2, 2022, doi: <https://doi.org/10.21070/pels.v2i2.1296>.
- [20] D. Figueiredo, G. Saraiva, J. Rebelo, R. Rodrigues, F. Cardoso, C. Wanzeller, P. Martins, and M. Abbasi, “Performance Evaluation Between HarperDB, Mongo DB and PostgreSQL,” in *Marketing and Smart Technologies. ICMarTech 2022. Smart Innovation, Systems and Technologies*, vol. 344. Springer, 2024. doi: [https://doi.org/10.1007/978-981-99-0333-7\\_7](https://doi.org/10.1007/978-981-99-0333-7_7).
- [21] M. Sals, A. Sghir, N. Rafalia, and J. Abouchabaka, “Analysis and comparison of NoSQL databases with relational databases: MongoDB and HBase versus MySQL,” *International Journal on Technical and Physical Problems of Engineering (IJTPE)*, vol. 55, no. 15, pp. 155–161, June 2023, ISSN 2077-3528.
- [22] J. Antas, R. R. Silva, and J. Bernardino, “Assessment of SQL and NoSQL Systems to Store and Mine COVID-19 Data,” *Computers*, vol. 11, no. 2, pp. 29, 2022. doi: <https://doi.org/10.3390/computers11020029>.
- [23] B. Alyasiri, B. Sahi, and N. AL-Khafaji, “NoSQL: Will it be an alternative to a relational database? MySQL vs MongoDB comparison,” in *Proceedings of 2nd International Multi-Disciplinary Conference Theme: Integrated Sciences and Technologies, IMDC-IST 2021*, Sakarya, EAI, 7-9 September 2021. doi: 10.4108/eai.7-9-2021.2314925.
- [24] M. Abukabar, S. Abukabar, and U. M. Bello, “Analyzing and Designing an Evaluation Benchmark for SQL and NoSQL Database Systems for Some Selected Higher Institution in Zamfara State,” *International Journal of Science for Global Sustainability (IJSGS)*, vol. 10, no. 2, p. 63, 2024, doi: 10.57233/ijsgs.v10i2.644.
- [25] H. Matallah, G. Belalem, and K. Bouamrane, “Comparative study between the MySQL relational database and the MongoDB NoSQL database,” *International Journal of Software Science and Computational Intelligence (IJSSCI)*, vol. 13, no. 3, pp. 38–63, 2021. doi: 10.4018/IJSSCI.2021070104.
- [26] C. Axell, E. Schøien, I. L. Thon, L. O. Vågene, and L. Tveiten, “Insertion speed of indexed spatial data: comparing MySQL, PostgreSQL and MongoDB,” 2022. [Online]: <https://folk.idi.ntnu.no/baf/eremcis/2022/Group02.pdf>.
- [27] M. Reichardt, M. Gundall, and H. D. Schotten, “Benchmarking the operation times of NoSQL and MySQL databases for Python clients,” in *IECON 2021 – 47th Annual Conference of the IEEE Industrial Electronics Society*, Toronto, Canada, 2021, pp. 1-8, doi: 10.1109/IECON48115.2021.9589382.
- [28] N. Naufal, S. Nurkhodijah, G. B. Anugrah, A. Pratama, M. I. Rabbani, F. A. Dilla, T. N. Anggraeni, and T. Firmansyah, “Comparisional analysis of MySQL and MongoDB response time query performance,” *Jurnal Informatika Dan Teknologi Komputer (JITEK)*, vol. 2, no. 2, pp. 158-166, 2022. <https://doi.org/10.55606/jitek.v2i2.245>
- [29] Citus Data Documentation, [Online]: <https://docs.citusdata.com>
- [30] L. F. Silva and J. V. F. Lima, “An evaluation of relational and NoSQL distributed databases on a low-power

- cluster,” *J. Supercomput.*, vol. 79, pp. 13402–13420, 2023. <https://doi.org/10.1007/s11227-023-05166-7>
- [31] U. Cubukcu, O. Erdogan, S. Pathak, S. Sannakkayala, and M. Slot, “Citrus: Distributed PostgreSQL for data-intensive applications,” in *Proc. of the 2021 International Conference on Management of Data*, ACM, 2021, doi: 10.1145/3448016.3457551.
- [32] MongoDB Documentation, [Online]: <https://www.mongodb.com/docs/manual/>
- [33] T. Taipalus, “Database management system performance comparisons: A systematic literature review,” *J. Syst. Softw.*, vol. 208, no. 111872, 2024, doi: 10.1016/j.jss.2023.111872.
- [34] Transaction Processing Council Benchmark H (Decision Support) Standard Specification, [Online]: [https://www.tpc.org/TPC\\_Documents\\_Current\\_Versions/pdf/TPC-H\\_v3.0.1.pdf](https://www.tpc.org/TPC_Documents_Current_Versions/pdf/TPC-H_v3.0.1.pdf).
- [35] R, “R: A Language and Environment for Statistical Computing,” R. C. Team, Producer, & R Foundation for Statistical Computing, R version 4.4.0, 2024. Available: <https://www.R-project.org>.
- [36] H. Wickham et al., “Welcome to the Tidyverse,” *J. Open-Source Softw.*, vol. 4, no. 43, pp. 1–6, 2019, doi: 10.21105/joss.01686.
- [37] I. Patil, “Visualizations with statistical details: The 'ggstatsplot' approach,” *J. Open Source Softw.*, vol. 6, no. 61, 2021, doi: 10.21105/joss.03167.
- [38] M. Kuhn and J. Silge, *Tidy Modeling with R*, Sebastopol, California, USA: O'Reilly, 2022.
- [39] P. Biecek, “Dalex: Explainers for complex predictive models in R,” *J. Mach. Learn. Res.*, vol. 19, no. 84, pp. 1–5, 2018.



**Marin FOTACHE** graduated from the Faculty of Economics at Alexandru Ioan Cuza University of Iasi, Romania in 1989. He holds a PhD diploma in Business Information Systems (Business Informatics) from 2000 and he had gone through all didactic positions since 1990 when he joined the staff of Al. I. Cuza University, from teaching assistant in 1990, to full professor in 2002. Currently he is professor within the Department of Accounting, Business Informatics and Statistics in the Faculty of Economics and Business Administration at

Alexandru Ioan Cuza University. He is the (co)author of books and journal/conference articles in areas such as SQL, database design, NoSQL, Big Data, Data Engineering and Machine Learning.



**Cătălina BADEA** completed her Master's degree in Data Mining at the Faculty of Economics and Business Administration, Alexandru Ioan Cuza University of Iași. Since 2024, she has been enrolled as a PhD student at the Doctoral School of Economics and Business Administration, focusing on Business Informatics. Her doctoral research approaches performance and architectural problems in Big Data platforms, exploring the transition from Data Lake to Data Lakehouse. Her conference participation includes RoEduNet: Networking in Education and Research 2024, Globalization

and Higher Education in Economics and Business Administration 2024, and International Conference on Informatics in Economy (IE 2025). She serves as an associate teaching staff

member at Alexandru Ioan Cuza University of Iași, where she conducts laboratory sessions in database systems. Her research interests focus on Big Data system architectures and the performance of data processing in business applications.



**Marius-Iulian CLUCI** graduated from the Faculty of Economics and Business Administration at Alexandru Ioan Cuza University of Iași, Romania. He holds a master's degree in Software Development and Business Information Systems and currently he is a Ph.D. candidate at the Doctoral School of Economics and Business Administration. His research focuses on Big Data systems, the integration of Machine Learning in Apache Spark, and performance benchmarking of modern data architectures. He has taught on topics related to Big Data, Machine Learning, OpenStack, and distributed computing. He works as a Cloud Engineer on Microsoft Azure, designing scalable data integration and processing solutions. His academic contributions include papers on TPC-H benchmarking, Spark optimizations, and schema evolution. His areas of expertise include cloud computing, Apache Spark, Machine Learning, and data integration.



**Codrin-Stefan Esanu** graduated from the Faculty of Economics and Business Administration Iasi, specializing in Business Informatics, and obtained his Master's degree in Software Development and Business Information Systems. Currently, he is pursuing a Ph.D. at the Doctoral School of Economics and Business Administration. Started his career in the IT Service Management industry at Capgemini, he advanced quickly towards management roles. Leveraging the solid foundation in managing critical IT services, he successfully transitioned to a technical DevOps role at Cegeka Romania. Additionally, he serves as a university associate lecturer at Alexandru Ioan Cuza University in Iași, teaching courses in databases, Big Data and distributed computing. His professional expertise includes Linux administration, Scripting, Infrastructure as Code (Puppet, Ansible), Containers orchestration (OpenShift), GitOps (ArgoCD) and database systems like PostgreSQL, Citus, and Neo4j.

## An Overview of Big Data and NoSQL in the Video Game Industry

Cristiana COSTAN

Bucharest University of Economic Studies

Faculty of Cybernetics, Statistics and Economic Informatics

Bucharest, Romania

[costancristiana21@stud.ase.ro](mailto:costancristiana21@stud.ase.ro)

*The connection between relational and non-relational databases and game development is covered in this article. The concept of big data and its implications in the gaming industry is also stated, the capability to store user data becoming essential for marketing and game improvements. The ideas in this article are reinforced by creating a web application to see how NoSQL databases can be implemented in game development and how they can help us extract relevant data about users. Overall, the findings of this paper are intended to guide people who want to be part of or are part of the gaming industry in the choices related to which database to utilize in future games and to provide insights on how they might obtain information about potential consumers to improve their products or services.*

**Keywords:** Big Data, NoSQL, Video Games, Gaming Industry

### 1 Introduction

The video game market is shifting so quickly that in order to have an impact, game developers need to design products with a strong emphasis on user needs. Companies are investing more money than ever in creating games designed to achieve popularity and generate profits. As video games have become a popular form of entertainment and relaxation, more and more individuals are choosing to spend their leisure time playing them. Game producers can offer potential consumers the experiences they desire due to the evolution of data and its integration into various industries. A lot of games have to be able to store information on player accounts, characters, or even in-game items. Depending on the situation, relational databases or non-relational databases may be used to store such information. For instance, non-relational databases are preferred when the stored data changes frequently, as their scalability is crucial in managing multiple modifications made in short periods of time. In contrast, relational databases are preferred when the game information is static and does not change often. Additionally, in some

cases, both types of databases can be utilized for the same product, allowing the game to include simultaneously static and dynamic components that can be stored in different formats, corresponding to each database used.

The volume of data generated by the large number of users who play video games is massive, therefore NoSQL is mostly utilized for storing such data because it is more scalable, faster, cheaper, and flexible, addressing challenges in real-time applications like online games that relational databases are unable to solve.

Big data is often used for analytics in the gaming industry. Analysis of player information assists in the development of marketing guidelines for the gaming industry and the improvement of games, whether they are still in testing or are already available on the market. Thus, gaming companies can effectively promote their game and retain as many of the users they attract.

### 2 Overview of Big Data

Big data refers to massive data sets with a large, complex, and varied structure [1]. These data are produced by online transactions, logs, posts, search queries,



mobile and computer applications. Their massive growth led to difficulties in storing, managing, analyzing, and visualizing data through standard database software tools.

There are three main characteristics that define big data: variety, velocity, and volume. Through variety, big data generally has three types, depending on the distinct sources it might come from structured, semi-structured, and unstructured. Structured data is defined as data that is already labeled and easily sorted, whereas unstructured data is unpredictable and challenging to analyze. Semi-structured data is in the middle, not conforming to specified fields but having tags to distinguish data items. The volume of big data is larger than terabytes and petabytes ( $10^{15}$  bytes), and its size still increases. Velocity, which is crucial for organizations when they acquire data, is the rate at which data may be created and processed.

According to Dr. Aditya Vailaya, an expert in computer science, machine learning, and statistical pattern recognition, large amounts of user information can be utilized to improve product recommendations or deliver targeted information [2]. Big data analytics has applications in merchandising, pricing, marketing, and advertising. Understanding users' interests based on previous purchases or preferences and understanding the characteristics of best-selling products are only two of the analytical uses of big data.

### 3 SQL vs NoSQL

Structured Query Language (SQL) is a non-procedural data access language developed in the 1970s at the IBM research laboratories [3]. Nowadays, most database management systems implement a dialect of SQL, setting SQL as the predominant language for enterprise databases in our surroundings. The term "NoSQL" was first utilized in

1998 to refer to a relational database that did not employ SQL [4]. The main reason for the NoSQL movement was stated by Eric Evans, a blogger who contributed to popularizing the term. He declared that this alternative appeared due to the need to solve problems that relational databases could not. Nowadays, as NoSQL might include SQL, it is usually considered as "not only SQL" by developers.

NoSQL databases are preferred over SQL databases in certain situations when scalability, high availability, and low latency are needed [5]. The scale-out strategy on which NoSQL databases are often based enables a cheaper process of scaling to large data volumes and provides the ability to upgrade or change the structure of a database with no downtime [6]. NoSQL is mostly utilized to manage big data and real-time web applications. Four different forms of data storage are included in NoSQL databases: document, key-value, graph, and column. The performance of NoSQL databases varies according to their data model.

Relational databases use the ACID (Atomicity, Consistency, Isolation, Durability) model for transaction processing due to its high reliability and consistency [7]. Atomicity refers to transaction failure and commits. If it were for a transaction to fail, the database would be left in the state before the transaction started. Only when all the transactions are successful the data would be updated. Consistency represents the stable state in which the database needs to be both before and after a transaction occurs. Isolation refers to a transaction being able to start only when another transaction finishes. Durability means that a completed transaction has permanent effects in case of later system failures. Because of their different priorities, NoSQL databases use the BASE (Basic Availability, Soft state, Eventual consistency) model, derived from the CAP (Consistency, Availability, and Partition-tolerance) theorem. Maintaining a single copy of current data, having the data available for updates, and being able to



continue operating even in the presence of network partitions represent the elements of the CAP theorem. “ACID properties are to RDBMS as BASE is to NoSQL systems.” [8]. Basic availability refers to the extent to which a user can modify the data. During periods of disconnection or synchronization delays, the user can still make changes and his actions will synchronize with the database later, once connectivity is restored. Soft state represents changing the state without any input, which ensures consistency. Eventual consistency means that if an updated database element is not further updated for a significant period of time, all users will see the same updated item value.

Addressing relational database management systems (RDBMS) vs NoSQL in detail, both have their advantages and disadvantages. NoSQL is easily scalable, the database does not necessarily need administrators, is faster, cheaper, and more flexible, some of the suppliers can manage hardware failure, and is utilized for big data applications. RDBMS is easy to design, execute, maintain, and use, and information is kept in one place, being secured. When looking at drawbacks, each disadvantage of NoSQL can be solved by SQL and vice-versa. For example, SQL presents issues in high availability, which can be managed by using NoSQL.

Modern big data-driven applications require the performance and scalability offered by NoSQL. Cisco replaced Oracle RAC with Neo4J, while Netflix switched from Oracle RDBMS to Apache Cassandra. These changes improved latency, shortened query times, and decreased expenses. These are only a few of the benefits brought by the usage of NoSQL databases. NoSQL databases are considered a better fit for handling large datasets with a lot of data streaming quickly from various sources [9].

#### **4 The Video Game Industry**

Video gaming is considered to belong to the field of culture and involves recreational activities [10], allowing individuals to distance themselves from their everyday problems by redirecting their attention to new goals from within imaginary worlds.

“The development of this industry characterizes our generation: fast paced, technologically oriented, and targeted toward the young and young at heart.” [11] The first computer game, called “Spacewar”, was developed more than 50 years ago by Steve Russell, a Massachusetts Institute of Technology (MIT) student, as a demonstration of expertise and not for profit. In a short period of time, this game gained popularity within the United States of America (USA) and entrepreneurs sought the opportunity to make a profit off of this, game development becoming a business. The starting point of this industry began from the enjoyment of Nolan Bushnell playing Spacewar during his time at university, which led to the development of “Atari” in 1972.

Nowadays, the video game industry is growing rapidly, with over 10 video game genres, including sandbox, real-time strategy (RTS), shooters, and multiplayer online battle arena (MOBA) [12]. The strong competition determines companies to constantly improve game graphics and features in order to enhance user experience. User feedback represents the key factor in game development. In 2018 the production costs of some video games ranged from \$100 million to \$500 million. These gigantic investments represent a risk taken by the companies, the profits depending solely on the users.

#### **5 SQL and NoSQL in the Gaming Industry**

Choosing RDBMS or NoSQL when creating a game is up to the developers nowadays. If the game is based on static elements one can utilize relational databases and whether a faster save of the data is needed NoSQL can be employed. The benefits of NoSQL can be

seen, for instance, in scoreboards and player statistics, which are dynamic components. Table 1 highlights the main

advantages of utilizing NoSQL databases in the development of video games.

**Table 1.** Advantages of NoSQL in the Video Game Industry

Use Case	Advantages of Utilizing NoSQL
Storing User Profiles	NoSQL's flexibility and scalability enable the rapid creation and continuous updates of user profiles, allowing many users to manage their accounts efficiently.
Storing User Items	In video games where inventories contain items with attributes that change based on user modifications, NoSQL can efficiently manage a wide variety of items with diverse attributes, handling large amounts of data.
Storing Game States Which Adapt to Player Choices	NoSQL efficiently stores and manages dynamic data from player decisions that impact game progression and outcomes.
Storing Leaderboard	NoSQL is ideal for managing leaderboard data, enabling fast updates and optimized querying.
Storing Player Statistics	Player statistics are frequently generated during gameplay, and NoSQL is crucial for handling the volumes of data produced.

In 2018 Fortnite announced hitting a new peak of 3.4 million concurrent players and shared within their article the “challenges of rapidly scaling a game and its online services” [13]. Fortnite has a service named MCP which is comprised of 9 MongoDB shards, utilized for retrieving data regarding game profiles, statistics, items, matchmaking info, and more.

In the same year, Riot Games wrote in their blog about deploying League of Legends to more than 12 disparate game shards, divided by regions [14]. MySQL was employed as the database backend for storing player account information, Riot putting a high priority on ACID. Both Riot and their parent company, Tencent, widely utilize MySQL and thus have a significant number of engineers who are familiar with how it operates. Because of this and the type of data that needs to be stored, MySQL is employed to protect user accounts.

In 2022 Amazon Web Services (AWS) posted in their AWS for Games Blog the architecture employed in “New World”, a massively multiplayer online role-playing

game [15]. Every second, millions of states change, thus appearing their need of utilizing Amazon DynamoDB, a NoSQL database which provides high performance and reliability and is suitable for game data, which frequently lacks organization and coherence. Moreover, employing Amazon DynamoDB allowed developers to successfully expand the game from a singular test server during development to over 500 in production without making any changes to the code.

Blizzard Entertainment, a video game developer and publisher, has listed on its careers page many jobs, including in Engineering & Technology. For Overwatch, a game which in 2024 surpassed 100 million players [16], experience with SQL databases was a plus, whilst for World of Warcraft (WoW), another popular game released by Blizzard, knowledge of relational or non-relational databases represented a plus. It is possible to conclude from these optional requirements that WoW uses both SQL and NoSQL, while Overwatch uses SQL.

Thus, the decision between RDBMS and NoSQL depends on the needs of the game developers. Developers prefer RDBMS

when dealing with static data, such as game characters, fixed-attribute items, game rules, and dialogues that don't change based on player actions. In contrast, NoSQL is more popular when dealing with player statistics, multiplayer leaderboards, match scores, and game states that adapt to player choices.

## **6 Big Data's Impact on Gaming**

The gaming industry typically utilizes big data for analytics. The term "big data analytics" refers to the processing and analysis of massive amounts of data in order to gain useful information [17]. With the aim to assist analysts in making data-informed decisions, big data analytics makes it possible to find trends, patterns, and correlations in large amounts of raw data.

Big data analytics offers those in the gaming sector the ability to analyze large data sets, consumer preferences, market trends, and other crucial business data [18]. This information aids gaming companies in improving player experiences and developing customized products and services. Furthermore, a new era of gaming has emerged as a result of the application and review of game analytics, in which data analysis is now crucial to the success of game development businesses rather than just advantageous.

Nowadays, various methods are employed in big data analytics to enhance the gaming experience and increase user engagement. Player behavior analysis, which examines in-game movements, purchase history, and playtime patterns can lead to improved engagement and higher customer loyalty. Predictive analytics utilize historical data to predict future trends and player actions, aiding in improving player satisfaction and reducing game dropouts. Personalization algorithms leverage machine learning to create game recommendations, in-game content customization, and adapt advertising messages based on player

information. Real-time analytics enable immediate correction of any problems, improving the overall gaming experience by tracking and examining server performance, live player behavior, and in-game transactions. Heat maps and spatial data analysis focus on player activities from within the game environment in order to improve level design, balance game difficulty, and ensure an engaging player experience. A/B testing allows game developers to compare two versions of a game or a feature among various player segments to determine which performs better and is essential for making data-driven decisions about game design and updates, ensuring that modifications reflect user preferences. Finally, the economic analysis examines player spending patterns, the balance of the virtual economy, and pricing strategies to optimize in-game economies, ensuring they remain engaging, fair, and profitable.

In the rapidly evolving gaming industry, these analytical methods provide gaming companies valuable insights on player behavior and market trends, enabling them to provide their customers with quality products and services that suit their preferences.

Ubisoft, SEGA Europe, and Kolibri Games represent only some examples from the gaming industry in which big data analytics are employed to ensure player engagement through personalized content, targeted marketing strategies, and gameplay adjustments. SEGA, for instance, gathers more than 25,000 data events each second, such as player behavior and in-game interactions.

Another interesting topic is related to how big data analytics can be improved with the utilization of machine learning (ML). Ubisoft leverages machine learning, which evolved from years of data gathering, to analyze player behaviors and enhance game features [19]. Since the early 2010s, advancements in tools and methods have allowed them to move beyond basic analysis to applying machine learning for actionable

insights and innovation.

The involvement of machine learning in behavioral analysis is also suggested in an article that discusses the use of neural networks to observe patterns and behaviors of players in the game World of Warcraft (WoW), a popular Massively Multiplayer Online Role-Playing Game (MMORPG) [20]. The characteristics utilized were analyzed using big data and data visualization tools. This experiment was conducted with the aim of showing how player analysis in an online gaming environment works on a small scale when using neural networks. Additionally, other analysis methods from those previously mentioned in the paper have been used for various studies regarding the behavior of WoW players. These demonstrations show us that by using big data and neural networks, companies can streamline their analyses and improve their games more easily.

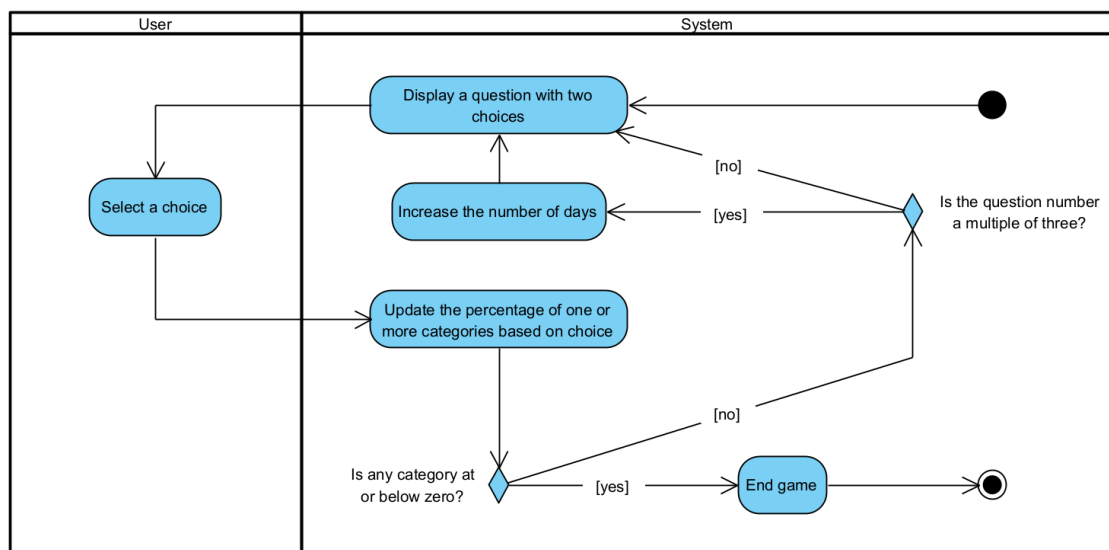
One type of artificial intelligence (AI) program that can detect and produce text, among other things, is called a large language model (LLM) [21]. Massive amounts of data are utilized to train LLMs, which are based on machine learning. Although there is still room for improvement in the understanding of data

by the LLMs, accurate outputs can be achieved when a more detailed and precise user input request is made [22]. This study also demonstrates the possibilities offered by the use of neural networks in the gaming industry, this time analyzing data related to the behavior of players in a gaming community, the data being extracted from an online forum for online board games. It has been discovered that exploratory data analysis (EDA) can be used effectively in this context, being an analysis that provides important insights from data.

These findings underscore the importance of big data in the gaming industry and highlight how game developers can benefit greatly from the analysis that can be performed with this data. Additionally, the chapter illustrates how machine learning has improved this sector by making it easier to understand players through big data analysis.

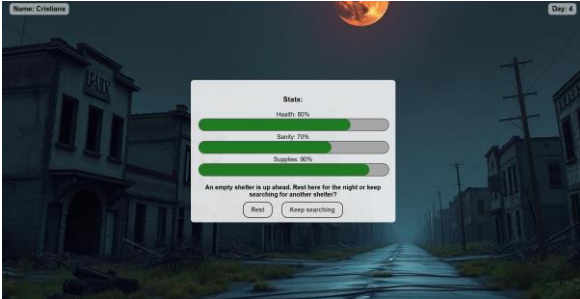
## 7 Building a Game

For a deeper understanding of NoSQL in the gaming industry I created a browser game with states that adapt to player decisions, as illustrated in Figure 1. It contains choices from three distinct categories: health, sanity, and supplies. On every in-game day, each user answers three questions.



**Fig. 1.** Activity Diagram for Game Flow

Each category's percentage is affected by the choices made, and the game is over when one of the categories hits zero. A visual representation of the game's core features is provided in Figure 2.



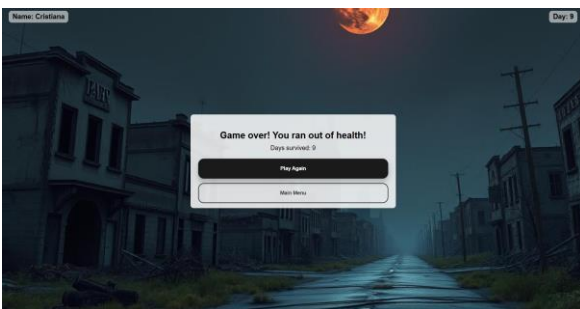
**Fig. 2.** Gameplay Overview

Since the game is still in its early stages, it merely keeps track of user data to generate the online leaderboard, which shows details about the players and how many days each of them survived. As depicted in Figure 3, every time the player completes the game, a POST request is made to the MongoDB database to insert the leaderboard data.

```
data = request.get_json()
days = int(data.get('days_survived'))
user = users_collection.find_one({'username': username})
high_score = int(user['high_score'])
if days < high_score:
    result = users_collection.update_one(
        {'username': username},
        {'$set': {
            'last_score': days
        }}
    )
else:
    result = users_collection.update_one(
        {'username': username},
        {'$set': {
            'last_score': days,
            'high_score': days
        }}
    )
```

**Fig. 3.** Python code for Updating User Score in MongoDB

The number of days survived is checked in the Python backend, entered into the database as the last score, or updated as both the last and high scores, and displayed on the final screen, as shown in Figure 4.



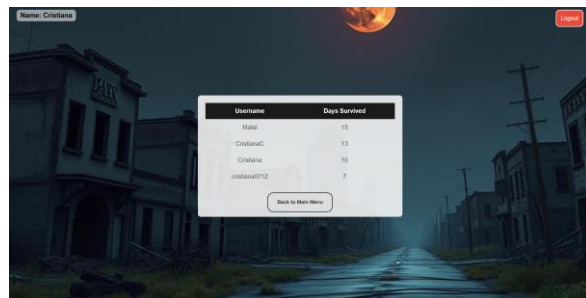
**Fig. 4.** Game Over Screen

All users with a score greater than zero days of surviving have their data extracted from the database and sorted in descending order, with a limit of ten users. This data is utilized to create the leaderboard. The code utilized for this process is displayed in Figure 5.

```
app.route('/leaderboard')
def leaderboard():
    if username not in session:
        return redirect(url_for('login'))
    username = session['username']
    leaderboard_info = users_collection.find({'high_score': {'$gt': 0}}).sort({'high_score': -1}).limit(10)
    return render_template('leaderboard.html', username=username, leaderboard_info=leaderboard_info)
```

**Fig. 5.** Python Code for Retrieving Leaderboard Data

If we were to think about the future of the game and the possibility of it becoming used by a multitude of people, the leaderboard data, as depicted in the leaderboard from Figure 6, would be extracted from a huge dataset, requiring NoSQL for its management. This will ensure continuous availability, low response times, and high-speed writes, ensuring that the database will intake data rapidly.



**Fig. 6.** Scoring Leaderboard

Taking all this into account, relational databases are recommended for developing a game that contains static features, including items with fixed basic attributes. When a game has dynamic elements, such as character traits, inventory items, states that change based on player choices, or player statistics, NoSQL databases are preferred over relational databases.

## 8 Conclusions and Future Work

The existence of big data has changed the gaming industry, and the emergence of NoSQL has made it easier to store this type

of data. Both relational and NoSQL databases can provide various benefits for game developers, and the choice of a database type is made depending on the needs of the respective game. NoSQL databases are preferred over relational databases in cases where games will have a huge number of players, each statistic of each player having to be stored in the database. Additionally, NoSQL makes it easier to store large amounts of data from which information can be collected, processed using pre-existing algorithms, and used to easily analyze player behavior. In terms of data storage, relational databases and NoSQL databases are both significant for the gaming sector. While NoSQL databases are more appropriate for handling dynamic and massive volumes of player data, relational databases are better suited when utilizing static elements.

Regarding the future development of the application, potential analyses can be conducted to improve the game based on the data provided by the players, such as time played and maximum scores achieved. As a concrete example, if a player takes a long time to reach an average score, the game can be altered and made simpler, or difficulty levels can be added that users can choose at the beginning of the game.

## References

- [1] S. Sagioglu and D. Sinanc, "Big data: A review," *2013 International Conference on Collaboration Technologies and Systems (CTS)*, San Diego, CA, USA, 2013, pp. 42-47.
- [2] N. Salim, "What's All the Buzz Around 'Big Data?': Meet Dr. Aditya Vailaya [The Good, the Bad, and the Ugly: Engineering FACTS]," *IEEE Women in Engineering Magazine*, vol. 6, no. 2, Dec 2012, pp. 24-31.
- [3] D. C. Kreines, "Oracle SQL: the essential reference," *O'Reilly Media, Inc.*, 2000.
- [4] C. Strauch, U. L. Sites and W. Kriha, "NoSQL databases," *Lecture Notes, Stuttgart Media University*, vol. 20, no. 24, Feb 2011.
- [5] B. Feltrin, "SQL versus NOSQL performance in Big Data analytics," *Master's thesis*, University of Zagreb, Faculty of Economics, Zagreb, 2022.
- [6] MongoDB, "When to Use a NoSQL Database," [mongodb.com. https://www.mongodb.com/resources/basics/databases/nosql-explained/when-to-use-nosql](https://www.mongodb.com/resources/basics/databases/nosql-explained/when-to-use-nosql) (accessed Dec 2024).
- [7] G. C. Deepak, "A critical comparison of NOSQL databases in the context of ACID and BASE," *Culminating Projects in Information Assurance*, St Cloud State University, 2016.
- [8] V. N. Gudivada, D. Rao and V. V. Raghavan, "NoSQL Systems for Big Data Management," *2014 IEEE World Congress on Services*, Anchorage, AK, USA, 2014, pp. 190-197.
- [9] J. Bhogal and I. Choksi, "Handling Big Data Using NoSQL," *2015 IEEE 29th International Conference on Advanced Information Networking and Applications Workshops*, Gwangju, Korea (South), 2015, pp. 393-398.
- [10] R. Baltezarevic, B. Baltezarevic and V. Baltezarevic, "The video gaming industry: From play to revenue," *International Review*, no. 3-4, 2018, pp. 71-76.
- [11] P. Zackariasson and T. L. Wilson, "Paradigm shifts in the video game industry," *Competitiveness Review*, vol. 20, no. 2, Dec 2012, pp. 139-151.
- [12] D. Pavlovic, "Video Game Genres: Everything You Need to Know," [hp.com. https://www.hp.com/us-en/shop/tech-takes/video-game-genres](https://www.hp.com/us-en/shop/tech-takes/video-game-genres) (accessed Dec 2024).
- [13] The Epic Team, "Postmortem of Service Outage at 3.4M CCU," [fortnite.com. https://www.fortnite.com/news/postmortem-of-service-outage-at-3-4m-ccu](https://www.fortnite.com/news/postmortem-of-service-outage-at-3-4m-ccu) (accessed Dec 2024).
- [14] Riot Games, "Globalizing Player Accounts," [technology.riotgames.com. https://technology.riotgames.com/news/](https://technology.riotgames.com/news/)

- globalizing-player-accounts (accessed Dec 2024).
- [15] AWS for Games Content Team and N. Walsh, “The Unique Architecture behind Amazon Games’ Seamless MMO New World,” [aws.amazon.com](https://aws.amazon.com/blogs/gametech/the-unique-architecture-behind-amazon-games-seamless-mmo-new-world/). <https://aws.amazon.com/blogs/gametech/the-unique-architecture-behind-amazon-games-seamless-mmo-new-world/> (accessed Dec 2024).
- [16] Blizzard Entertainment, “Relive Your Overwatch Glory!,” [overwatch.blizzard.com](https://overwatch.blizzard.com/en-us/news/24104603/relive-your-overwatch-glory/). <https://overwatch.blizzard.com/en-us/news/24104603/relive-your-overwatch-glory/> (accessed Dec 2024).
- [17] T. Mucci and C. Stryker, “What is big data analytics?,” [ibm.com](https://www.ibm.com/think/topics/big-data-analytics). <https://www.ibm.com/think/topics/big-data-analytics> (accessed Dec 2024).
- [18] F. Hagverdiyev, “Data Governance in Gaming Industry,” *Problems of Information Technology*, vol. 15, no. 1, 2024, pp. 44-51.
- [19] C. Watters, “Shaping the Future of Games: How Machine Learning and Data Science Will Help Transform the Industry,” [news.ubisoft.com](https://news.ubisoft.com/en-us/article/5eZ3g9FHSllkSlkp12lyVL/shaping-the-future-of-games-how-machine-learning-and-data-science-will-help-transform-the-industry). <https://news.ubisoft.com/en-us/article/5eZ3g9FHSllkSlkp12lyVL/shaping-the-future-of-games-how-machine-learning-and-data-science-will-help-transform-the-industry> (accessed Dec 2024).
- [20] V. P. Barros and P. Notargiacomo, “Big data analytics in cloud gaming: Players’ patterns recognition using artificial neural networks,” *2016 IEEE International Conference on Big Data (Big Data)*, Washington, DC, USA, 2016, pp. 1680-1689.
- [21] Cloudflare, “What is a large language model (LLM)?,” [cloudflare.com](https://www.cloudflare.com/learning/ai/what-is-large-language-model/). <https://www.cloudflare.com/learning/ai/what-is-large-language-model/> (accessed Dec 2024).
- [22] V. Sharma and M. Dave, “SQL and NoSQL Databases,” *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 2, no. 8, August 2012.



**Cristiana COSTAN** graduated from the Faculty of Cybernetics, Statistics and Economic Informatics of the Academy of Economic Studies in 2024, obtaining a Bachelor’s Degree in Economic Informatics. She is currently pursuing a master’s degree in Databases - Business Support and will earn her degree in 2026. Her main areas of interest include data analysis, database management, and web development.



# Evaluating Deep Learning and Machine Learning Models in Federated Learning for Credit Card Fraud Detection. A comparative study

Şener ALI

Bucharest University of Economic Studies

Faculty of Cybernetics, Statistics and Economic Informatics

Bucharest, Romania

[alisener20@stud.ase.ro](mailto:alisener20@stud.ase.ro)

*This study aims to evaluate the effectiveness of both machine learning and deep learning algorithms for credit card fraud detection in the context of a federated learning framework. The fast evolution of digital banking created better experiences for customers and facilitated their access to financial services, but it has simultaneously opened new pathways for cybercriminals, making real-time fraud detection essential. Both models used in this study, XGBoost and a neural network, were trained on a publicly available dataset containing highly imbalanced data, reflecting realistic fraud scenarios. Results demonstrate that both models achieved high accuracy, yet the neural network consistently outperformed XGBoost across critical metrics such as precision, recall, and F1 score. This indicates a superior ability of deep learning models to detect fraudulent transactions in federated learning environments, highlighting their potential to improve financial security through collaborative yet privacy-preserving approaches.*

**Keywords:** Federated Learning, Credit Card Fraud Detection, Privacy Protection, Machine Learning, Deep Learning

## 1 Introduction

In the last years, digital banking has evolved significantly, but this progress has also given rise to more advanced fraud techniques. Cybercriminals can use advanced methods, such as stealing card information used in online transactions, to commit financial crimes. Besides the traditional ways, such as card theft, criminals have at this moment online methods to commit fraud [1].

Based on the central bank's reports, billions are lost every year due to fraudulent activities. Besides the financial losses of the clients, these frauds can have a negative impact on the trust of people in financial institutions. Undermining the credibility of the banking system can create crashes on financial markets and threaten the global economy [2].

To keep a high level of trust in financial institutions, we can employ two mechanisms: fraud prevention and fraud detection. In this study, we will focus on fraud detection, the mechanism that tries to

detect fraudulent transactions in real time, through the use of federated learning (FL).

Introduced by Google in 2017 to improve their keyboard text prediction, federated learning became an approach used for training AI models without the need of sharing the data across institutions [3]. Unlike the traditional approaches where aggregating the data in a centralized dataset was necessary, the federated learning offers a decentralized approach. Each institution trains a model locally on its dataset and sends it to the server, where it is aggregated to the global model. This approach enables us to obtain a model trained on all the available data while the data remains confidential [4].

The key point in developing efficient AI models is to have diverse and high-quality data. The obligation to keep the data confidential presents a significant challenge for financial institutions because they can not share their data to create a centralized dataset and train an AI model on it [5]. Moreover, based on the dataset of



an institution, the AI model is trained to detect only the frauds that the bank has already experienced, but it can remain susceptible to the attacks that other banks encountered. This makes the entire financial system more vulnerable, allowing cybercriminals to repeatedly exploit the same type of vulnerability across multiple financial institutions [6].

To address the challenges of the centralized approach, FL solves these limitations by offering the financial institutions the possibility to train a global model without sharing their data. Each bank trains a local model on its own dataset. After training, the model updates are sent to a central server and aggregated to the global model [7]. Adopting this approach, the financial institutions solve both problems, they train a model on all their available data while keeping them confidential. By leveraging knowledge from multiple institutions, FL strengthens the entire financial system by protecting the companies even from fraud cases they have not previously experienced [8].

The application of FL in fraud detection has gained significant attention due to the growing sophistication of fraudulent activities. Traditional machine learning (ML) and deep learning (DL) models have been widely used in fraud detection, but their performance in a federated setting remains an area of active research. While deep learning models, such as convolutional neural networks and recurrent neural networks, can capture complex fraud patterns, traditional ML models, such as decision trees and support vector machines, may offer advantages in interpretability and computational efficiency [9].

Given the increasing adoption of FL and the diverse range of ML and DL models, it is crucial to evaluate their effectiveness in a federated setting. This study aims to investigate the comparative performance of these models in detecting credit card fraud. Specifically, the research seeks to answer to the following question:

**RQ.** Which performs better in a Federated Learning framework for credit card fraud detection: a machine learning model or a deep learning model?

By addressing this research question, this study aims to provide a comprehensive evaluation of ML and DL models in FL for fraud detection, offering insights into their real-world applicability and potential for improving financial security.

The rest of the paper is organized as follows. In Section 2, related work is discussed. Section 3 provides an analysis of the dataset. Section 4 gives the details of the federated learning fraud detection implementation. The results of the study and their interpretation are presented in section 5. Conclusions are discussed in section 6.

## 2 Literature review

The importance of credit card fraud detection is represented by the number of public available works. There are numerous articles that talk about this topic or researches that are trying to find the best method to stop fraudsters.

Given the new context of the rise of online financial transactions and the improvements of AI in the last years, the machine learning algorithms have gotten more attention to become a relevant solution for fraud detection [10]. One of the algorithms used for handling the complexity of credit card problems is XGBoost [11]. Being one of the gradient boosting frameworks, the algorithm has a focus on regularization and adopts an iterative approach, improving its accuracy over time.

Comparing XGBoost with other machine learning algorithms when it comes to detecting credit card fraud detection shows us that the XGBoost has better performance, obtaining better results for indicators such as accuracy, precision and recall [12].

Credit card fraud detection has two major challenges: the very limited time span in

which a transaction has to be accepted or rejected and the huge amount of transactions. Only VISA has millions of transactions every day. To address these challenges, the researchers thought that neural networks would be a great solution [13]. As shown in [14], using neural networks can be a great approach to detect credit card fraud.

Besides the two previous challenges we discussed, there is another challenge when it comes to fraud detection. The financial institutions can not share the data, it would be against the data privacy and safety policies [15]. As a result, each financial institution can train models that detect only the patterns of the attacks they experienced, not the patterns of the attacks experienced by other institutions. In this way, the fraudsters can attack each financial institution with the same pattern. What if we could prevent the other financial institutions from experiencing the same type of attack that other banks already experienced? To address this problem, we can implement federated learning.

Federated learning offers the possibility to train models on local devices and share just the model's updates. The main idea of federated learning is to train a global model on all the available data without sharing it. The local trained models are sent to the central server that aggregates the weights. After update, the global model is sent back to the local devices [16]. In this way, the global model's training takes advantage of all available data while keeping it confidential [17].

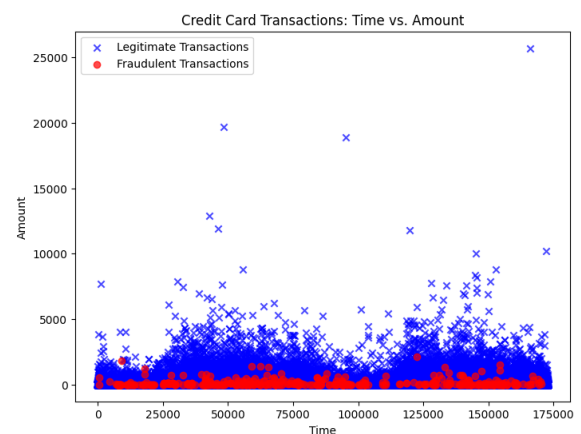
Given the three problems: the necessity for fast decisions, the need to handle large amount of transactions and the requirement to keep data confidential and the good results obtained in the other studies by both XGBoost and neural networks models, this study aims to compare the performance of these two algorithms in a federated learning architecture.

### 3 Data

For this study, we used a publicly available dataset that consists of 284,807 transactions [18]. Given the sensitivity of financial data, each transaction has 30 numerical features that have been transformed using Principal Component Analysis (PCA) to protect customers confidentiality.

Additionally, the transactions have two explicit features: Time, representing the seconds elapsed since the first transaction and Amount, denoting the transaction's value in monetary units. The dataset also contains a Class column that identifies whether a transaction is fraudulent (1) or legitimate (0).

As all financial frauds datasets, the dataset we used in this study is highly unbalanced. The fraudulent transactions make up only 0.1727% of the total data, with 492 fraud cases out of 284,807 transactions. To illustrate this imbalance, we created a 2D scatter plot where each point represents a transaction, colored by its class label. Legitimate transactions (blue) form the majority of the data points, while fraudulent transactions (red) are less frequent, reflecting the small proportion of fraud cases.



**Fig. 1.** Dataset visualization

For conducting this study, we needed to split the original dataset into five different smaller datasets. These subsets did not originate from actual banks, but were created to simulate local datasets. For

training each local model, 80% of the corresponding dataset was used as a training set, while the remaining 20% was reserved for testing. The initial dataset was used to test the global model.

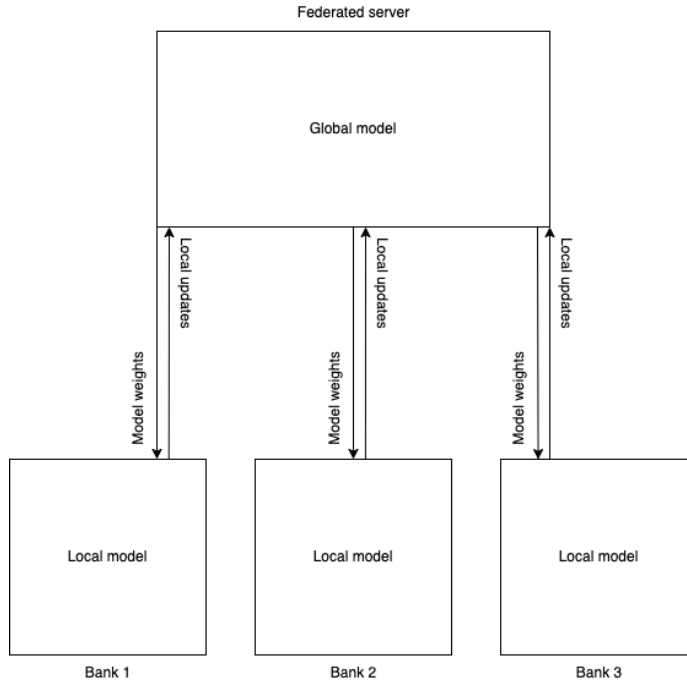
#### 4 Methodology

In this section, we explain the methodology employed in this research. We describe the technical approach, the algorithms implemented to address the research problem. Specifically, the methodology revolves around the

implementation of XGBoost and neural networks models within the Federated Learning framework.

##### A. Federated Learning architecture

Fig. 2 illustrates the FL architecture used in this research. Each client node operates independently with its own dataset, representing a financial institution (e.g., a bank). Once a local model is trained on a bank's dataset, the learned parameters (weights) are transmitted to a central server.



**Fig. 2.** Federated Learning Architecture

At the central server, the local updates are aggregated to improve the global model. The aggregation process involves computing an average of the local weights, where each bank's contribution is scaled by the number of samples it used during training. The aggregation is expressed by the formula [5]:

$$\omega_t = \sum_{k=1}^K \frac{n_k}{n} \omega_{t+1}^k \quad (1)$$

This approach ensures that the global model benefits from diverse data sources while preserving the privacy of each client's data. Before any updates are received, the server initializes the global model either with random parameters or, as

in our approach, by pretraining on an available dataset. Once the initial global model is distributed to the clients, each institution trains locally and sends back only its updated model parameters.

On the client side, each institution is responsible to train a local model on its own dataset and send the model weights to the server. This process is important because only the weights are shared, the data remains private.

##### B. XGBoost Model

The machine learning model chosen for this study was XGBoost. Each financial institution trains locally a model using the

XGBoost classifier, a gradient boosting framework effective with tabular data and robust in handling class imbalance. To ensure the optimal selection of hyperparameters, we integrated GridSearchCV. The grid included variations in tree depth, learning rates and number of estimators. In this way, different combinations of these parameters were evaluated based on their performance, measured by F1-score. This exhaustive search ensured that the selected hyperparameters were the best suited for handling the class imbalance challenge common in financial datasets.

### C. Neural Network Model

Our research employs a neural network model, a model well-suited for handling the imbalanced nature of financial frauds datasets.

The network consists of multiple layers:

- a. **Input Layer:** The input layer has two major responsibilities: to receive the data and to properly format its features, ensuring that the next layers receive standardized inputs. The job of the input layer is critical because it builds the foundation for accurate pattern recognition throughout the network.
- b. **Hidden Layers:** The hidden layers are the core of the network. The learning happens here. Our neural network implements three hidden layers, each with a progressively reduced number of neurons, from 128 neurons in the first layer down to 32 neurons in the final layer. Each hidden layer is followed by batch normalization, which improves the network's training stability, and ReLU, for improving the catch of complex patterns in the data. Moreover, we apply dropout regularization at a rate of 50% after each hidden layer to avoid the risk of overfitting.

- c. **Output Layer:** The output layer is responsible for preparing the data for getting out of the network. The output layer's job is to synthesize the information learned by the hidden layers into a final decision. The final input produces a logit, the raw output value, corresponding to the log-odds of the positive class. To get a probability score ranging between 0 and 1, the logit is passed through a sigmoid function. The probability score decides if the transaction is classified as fraudulent or legitimate.

To handle one of the biggest challenges of financial fraud detection, the highly imbalanced nature of datasets, we integrated a focal loss function. By doing this, the model can concentrate more on the minority class, the fraudulent cases, rather than being overly influenced by the majority class. In this way, the overall performance of the model is improved by learning more effectively from the minority class. Mathematically, the focal loss function is expressed as [19]:

$$FL(p_t) = -\alpha_t (1 - p_t)^\gamma \log(p_t) \quad (2)$$

## 5 Results and discussions

This section presents the results of our study. The discussion about each model's performance revolves around five key metrics: accuracy, precision, recall, F1 score and AUC-PR [20]. Moreover, for a better evaluation, we interpreted the confusion matrices.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

$$F1 \text{ score} = \frac{2*Precision*Recall}{Precision+Recall} \quad (6)$$

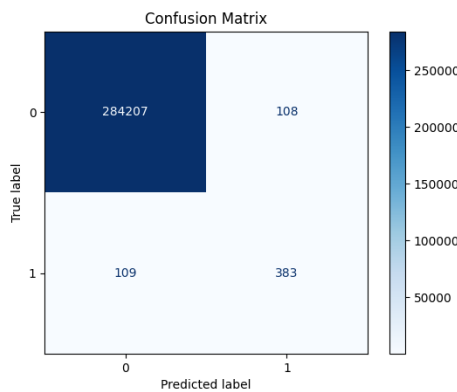
Method	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	AUC-PR (%)
XGBoost	99.91	73.88	73.58	73.73	71.88
Neural Networks	99.94	85.13	80.28	82.64	76.08

**Fig. 3. Results**

Fig. 3 presents the results obtained by each model in the federated learning framework. The results and its interpretations for the XGBoost and neural network models are discussed individually below.

#### A. XGBoost model

The high accuracy of 0.9991 achieved by the XGBoost model indicates that almost all predictions were correct. The percent of positive predictions is around 74% indicated by the precision of 0.7388. This also shows us that the percent of false positives is pretty high, around 26%. The recall value of 0.7358 obtained by this study shows us that 74% of the true positive instances were detected, while 26% of them went undetected. The F1 score of 0.7373 shows us the trade-off between precision and recall, highlighting both the model's strength in overall prediction accuracy and its limitations in precisely capturing all relevant cases. The model obtained an AUC-PR score of 0.7188. This suggests that it has a moderate balance between precision and recall across different thresholds.

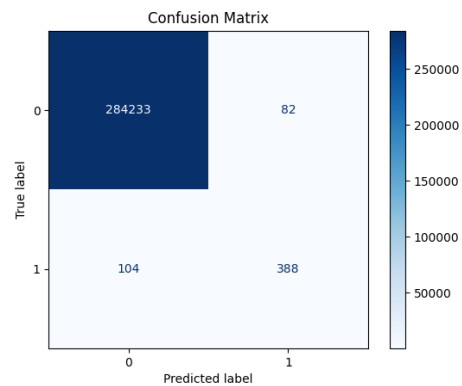
**Fig. 4. XGBoost - confusion matrix**

The confusion matrix presented in Fig. 4 illustrates the number of instances classified by the XGBoost model as true

negatives, false negatives, true positives and false positives. The model performed well in detecting non fraudulent transactions, with 284.207 instances classified as true negatives and only 108 false negatives. For positive cases, the model demonstrated worse performance, with 383 true positives, but 109 false positives. This indicates that around 22% of fraudulent transactions were not correctly identified by the model.

#### B. Neural Network Model

As shown in Fig. 3, the Neural Network model's accuracy has a value of 0.9994, meaning that nearly all predictions were correct. The precision of 0.8513 indicates that approximately 85% of positive predictions were correct, while the the rest of 15% were false positives. Similarly, the recall of 0.8028 means that the model correctly identified about 80% of the true positive instances, indicating that a portion of actual positives went undetected. The F1 score of 0.8264, which balances precision and recall, illustrates this trade-off, underscoring the model's effectiveness in overall prediction while also highlighting its limitations in accurately identifying all relevant instances. The AUC-PR of 0.7608 highlights that the Neural Network model achieves a stronger level of precision across various recall thresholds than the XGBoost model.

**Fig. 5. Neural Network - confusion matrix**

The confusion matrix represented by Fig. 5 shows the results obtained by the neural

network model. This matrix shows 284,233 correctly classified negative cases and 388 correctly identified positive cases. Compared to the XGBoost model, the false positives have decreased to 82, suggesting the model is more precise in avoiding incorrect positive classifications. Similarly, false negatives have reduced slightly to 104, enhancing the model's sensitivity by detecting a higher proportion of actual positives.

The results of this comparative study can serve as a starting point for developing a real-world fraud detection system based on the federated learning framework. Given its better performance, the neural network can be the primary candidate model for building such a system. The main limitation encountered during this study was the lack of sufficient data, as the confidentiality requirements restrict the availability of public financial datasets. Having multiple distinct datasets, rather than dividing one into several subsets, would likely have led to improved results. Future research can focus on extending the federated learning approach to other types of financial fraud, including insurance fraud, loan fraud or money laundering. Such studies would validate the applicability and effectiveness of federated learning framework across a broader spectrum of financial crime detection scenarios.

## 6. Conclusions

In this study we evaluated the performance of both XGBoost and neural network models within a federated learning model. In terms of accuracy, both models performed exceptionally. In key performance metrics such as precision, accuracy, F1 score, and AUC-PR score the neural network model outperformed the XGBoost model. While in previous studies, XGBoost had the best performance when compared to other machine learning models, the deep learning model demonstrated a better ability in detecting fraudulent transactions, making it a

promising choice for real-world cases. Overall, this study highlights the potential of federated learning to improve collaboration among financial institutions to develop better fraud detection systems.

## References

- [1] K. S. Amit, "An Overview of Digital Payment Frauds: Causes, Consequences, and Countermeasures," *Journal of Informatics Education and Research*, vol. 5, no. 1, pp. 2297-2311, 2025.
- [2] Yukun, "The Impact of Financial Fraud on Economic Stability: An Extensive Economic Analysis," *Journal of Internet Banking and Commerce*, vol. 28, no. 4, 2023.
- [3] Y. Wensi, Z. Yuhang, Y. Kejiang, L. Li and X. Cheng-Zhong, "FFD: A Federated Learning Based Method for Credit Card Fraud Detection," *Big Data – BigData 2019*, vol. 11514, pp. 21-31, 2019.
- [4] M. M. Priyanka, "Federated Learning: Opportunities and Challenges," in *Proceedings of ACM Conference (Conference'17)*, New York, 2021.
- [5] M. Brendan, M. Eider, R. Daniel, H. Seth and A. y. A. Balise, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, Florida, 2017.
- [6] M. Prakash, "Federated Learning for Cross-Bank Fraud Defense," *Intenational Journal of Computer Engineering & Technology*, vol. 16, no. 2, pp. 221-241, 2025.
- [7] G. Poonam, H. Jayesh and K. Prakash, "Federated Learning in Banking: A Secure and Intelligent Approach to Financial Fraud Detection," *Journal of Emerging Technologies and Innovative Research*, vol. 12, no. 6, pp. 629-636, 2025.
- [8] T. Nguyen, S. Kai, W. Siyao, G. Florian and G. Yike, "Privacy



- preservation in federated learning: An insightful survey from the GDPR perspective," *Computers & Security*, vol. 110, pp. 1-23, 2021.
- [9] Z. Han, "Federated Learning-Based Credit Card Fraud Detection: A Comparative Analysis of Advanced Machine Learning Models," *ITM Web of Conferences*, vol. 70, pp. 1-6, 2025.
- [10] A. Abdulalem, A. R. Shukor, H. O. Siti, A. E. E. Taiseer, A.-D. Arafat, N. Maged, E. Tusneem, E. Hashim and S. Abdu, "Financial Fraud Detection Based on Machine Learning: A Systematic Literature Review," *Machine Learning for Cybersecurity Threats, Challenges, and Opportunities II*, pp. 1-24, 26 September 2022.
- [11] R. N. Teuku, M. I. Ghalieb, M. Aga, H. Irsan, S. R. Edi and I. Rinaldi, "Credit Card Fraud Detection for Contemporary Financial Management Using XGBoost-Driven Machine Learning and Data Augmentation Techniques," *Indatu Journal of Management*, vol. 1, no. 1, pp. 29-35, 2023.
- [12] K. M. Krishna, Z. K. Mohammad and I. Ajay, "Credit Card Fraud Prediction Using XGBoost: An Ensemble Learning Approach," *International Journal of Information Retrieval Research (IJIRR)*, vol. 12, no. 2, pp. 1-17, 2022.
- [13] P. Raghavendra and S. Lokesh, "Credit Card Fraud Detection Using Neural Network," *International Journal of Students' Research in Technology & Management*, vol. 2, no. 2, pp. 84-88, 2015.
- [14] R. Asha and K. K. Suresh, "Credit card fraud detection using artificial neural network," *Global Transitions Proceedings*, vol. 2, no. 1, pp. 35-41, 2021.
- [15] K. A. Saif, J. A. Saif, D. Hussain and A. K. Muhammad, "Secure and Transparent Banking: Explainable AI-Driven Federated Learning Model for Financial Fraud Detection," *Journal of Risk and Financial Management*, vol. 18, no. 4, pp. 1-26, 2025.
- [16] Y. Qiang, L. Yang, C. Tianjian and T. Yongxin, "Federated Machine Learning: Concept and Applications," *ACM Transactions on Intelligent Systems and Technology*, vol. 10, no. 2, pp. 1-19, 2019.
- [17] Tomisin, M. S. Raj and P. Bernardi, "Transparency and Privacy: The Role of Explainable AI and Federated Learning in Financial Fraud Detection," *IEEE Access*, vol. 12, pp. 64551-64560, 2024.
- [18] M. L. G. -. ULB, "Credit Card Fraud Detection Dataset," Kaggle, 2018. [Online]. Available: <https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud>. [Accessed 4 February 2025].
- [19] L. Tsung-Yi, G. Priya, G. Ross, H. Kaiming and D. Piotr, "Focal Loss for Dense Object Detection," *IEEE International Conference on Computer Vision (ICCV)*, pp. 2999-3007, 2017.
- [20] N. M.Z and A. Amir, "Insights into Performance Fitness and Error Metrics for Machine Learning," *Architecture, Structures and Construction*, pp. 1-19, 2020.
- [21] M. Samaneh, B. Ali, S. Sima and F. Francesco, "Balancing privacy and performance in federated learning: A systematic literature review on methods and metrics," *Journal of Parallel and Distributed Computing*, vol. 192, 2024.



**Sener Ali** is a master's student at the Academy of Economic Studies in Bucharest, where he is expected to graduate in 2025. He currently works as a web developer, and his main areas of interest include web development, artificial intelligence, and blockchain.



## The Advantage of NoSQL Databases over SQL Databases

Mario-Tudor CHIRIAC

Bucharest University of Economic Studies

Faculty of Cybernetics, Statistics and Economic Informatics

Bucharest, Romania

[chiriadmario21@stud.ase.ro](mailto:chiriadmario21@stud.ase.ro)

*The aim of this article is to show the advantages of No-SQL databases compared to SQL (relational) databases. Traditionally, the concept of database is implicitly associated with relational databases, most often ignoring the huge potential that non-relational databases have in the fields such as BigData, Data Analysis, Artificial Intelligence. Besides, non-relational databases are also remarkable for their structure flexibility, their speed and for the distribution of their data on multiple servers simultaneously, for their fast writing and reading speed, which is crucial when analysing very large volumes of data.*

**Keywords:** Relational databases, non-relational databases, No-SQL databases, SQL databases, SQL, No-SQL, redundancy, efficiency, scalability, portability, large volume of data, big data, management of large volumes of data, big data flow.

### 1 Introduction

Before the invention of computers and electronic devices, information would be written on paper, stone, or tree bark. These methods are quite inefficient compared to the electronic storage, as they involve a high human presence in the process of writing and managing the information. Besides, the risks associated with these methods are high: damage to materials (paper, stone, wood) leading to information losses. It is almost impossible to recover the information unless there are backup copies. Duplicating these sources involves human resources and the process is very time-consuming.

An important step in the field of information storage was made with the invention of the magnetic tape in the 1960s. Such a magnetic tape would be read sequentially, which meant that in order to find the desired information, the entire magnetic tape had to be read. Compared to today's standards, this is outdated and inefficient.

The emergence of the Oracle corporation in 1977 facilitated the development of the Oracle Database relational database, which is this corporation's most well-known and

highly appreciated product. At the end of the ninth decade of the last century (more precisely, in 1998), the notion of a "No-SQL" database was also postulated, a term coined by Carlo Strozzi [1]. Initially, this term designated those relational databases that would not use SQL as a data manipulation language. Later, with the emergence of the Cassandra and Voldemort databases in the early 2010s, this term became an umbrella name for all non-relational databases.

The invention of the personal electronic computer – Altair 8800 in the 1970s, would mark yet another important stage in the field of technology development and information management, while also being an important moment in the history of technology. This was an essential step, because it facilitated the emergence of several computer models in the following years, such as the Apple II and the IBM PC (Model 5150). The year 1995 marked the launch of the famous Windows 95 operating system, which became a successful brand of the Microsoft corporation. The Windows operating system has become one of the most important operating systems in the whole world, being used everywhere, from

personal computers to servers within production processes.

The exponential evolution of technology, its availability at the highest level and the fast digitalization have enhanced the population's perception of how information is used and stored, something that can be seen today.

Below, I will analyse some fundamental concepts that can be found in the above-mentioned databases, listing some of the studies that made a comparative analysis of these aspects.

The practical part gives the opportunity to understand some key concepts, useful in the presentation of the notions to be discussed, such as: scalability, performance, diversity of query languages.

## **2. Literature review**

According to the specialised literature, we could say that No-SQL databases have several advantages over classic relational databases. No-SQL databases are designed with high scalability, which is essential nowadays, considering the high Internet access rate, as it is necessary for fast data processing. Instead of needing new servers, NoSQL databases can be scaled using the existing hardware. Thus, they can handle traffic growth and respond to demand without activity interruptions. Through this scaling, NoSQL databases become more robust and more performant, thus becoming a preferred choice for expanding data volumes. [2], [15]

In terms of performance, No-SQL databases work optimally even on devices equipped with a common hardware configuration, while relational databases require complex configurations, being specific to devices that are not readily available to the general public (Oracle Exadata) [3]. Thus, it could be said that non-relational databases are very useful for software developers and smaller businesses that do not have complex hardware resources, as they can fit on modestly-

configured devices.

The data manipulation language within relational databases is quite orderly and centralised, which applies to most relational databases (Oracle Database, MySQL, MariaDB, SQLite). The differences in syntax/structure are minimal for general queries, and they occur only in advanced functionalities such as multimedia, geographical facilities, etc. No-SQL databases provide a very high flexibility. There is no well-defined standard for the language syntax, as it varies from one model to another.

One of the most well-known languages is the MongoDB Query Language (MQL) used for MongoDB document-oriented databases [4]. This decentralised approach to query languages provides several advantages for both developers and programmers. One advantage is a better cooperation between programmers and developers, thus, a continuous flow of feedback being ensured, on functionalities, bugs, and potential improvements. Besides No-SQL databases are specifically designed to be able to manage huge amounts of data efficiently, unlike the classic relational models, where their rigorous structure and integrity restrictions make such processing impossible. More detailed information about No-SQL databases, and their use in the field of big data, shall be developed in the future chapters.

According to the article [3], in the US, an SQL developer earns \$84,328 per year, in average, while a No-SQL developer earns \$72,174 per year. Economically speaking, the SQL developer earns more, but the No-SQL developer has the advantage of flexibility provided by the vast scope of his knowledge. A No-SQL database expert can also work as a data analyst, where the average wages are much higher than those of a simple SQL or No-SQL developer [5].

### 3 Theoretical concepts related to SQL and No-SQL databases. Introduction to BigData

With the invention of the computer and the Internet, the world has experienced an irreversible and revolutionary transformation of the framework in which the development of our personal lives, economic processes and activities take place. The symbiosis between the Internet and the computer has triggered the acceleration of economic and cultural globalisation. The need for rapid and instant access to information and data is the basis of today's notions and fields of activity, such as: data science, data analysis, data mining, etc.

All these sciences form the basis of the so-called BigData. It is obvious that data has a physical support on which it is stored and, via computer networks, it is transmitted between a computer and several other computers, thus forming a network.

Therefore, a database is formation of one or more interdependent data collections, accompanied by a file describing, globally, the data and the links between them. [6],[16]. Nowadays, the most common and frequently used types of databases are: SQL (Standard Query Language) and No-SQL (Not Only Standard Query Language).

There are several well-established manufacturers, on the relational (SQL) and No-SQL databases, providing database management software services. The role of database management systems is of particular importance, because, such a tool allows us to take care of the administrative sides of a database (data backup and recovery, access rights management, database performance analysis) and its security (data access authorisation and control, data encryption, protection against external attacks). According to the ranking made by the db-engines.com[7] website, the first five positions are held by the following database systems: Oracle, MySQL, Microsoft SQL Server,

PostgreSQL and MongoDB.

The difference between the two types of data consists of several criteria, such as: *structural criteria*: redundancy, internal data organisation structure, query language, etc. The second criterion used to differentiate the two types of databases is related to *characteristics*: efficiency, allocated memory, speed of data processing, etc.

Based on these criteria, we can make a comparison between the two types of databases: The SQL model has a rigorous and strict structure, where redundancy is minimal and controlled. The form of organisation within an SQL-type database is represented by tables, where the following relationships can exist: *1:1 – one to one*, *1:m – many to many* and *m:m – many to many*.

All these relationships within a relational database create the database's logical diagram. In other words, the logical diagram represents the theoretical structure of the way in which data is organised and linked together.

Besides, the logical diagram also shows the complexity of a relational database, but there are also other important concepts in the field of relational databases, such as: primary key, foreign key, integrity constraints or form of normalisation.

Unlike the SQL model, the No-SQL model is more permissive from a structural and functional point of view. No-SQL databases are divided into 4 categories: document-oriented databases (MongoDB), column-oriented databases (Apache Cassandra), graph-oriented databases (Neo4j) and key-value pair-oriented databases (Redis). No-SQL databases have a wide range of uses and applications: from the e-commerce field (electronic commerce) to the BigData field, which will be addressed in this chapter. Other domains using these databases include social media, mobile apps, gaming, and streaming services. Examples of corporations using such databases include

Netflix, Uber, and Airbnb.[8]

Netflix uses No-SQL databases to store the huge amounts of customer-generated data (viewing history, customer personal data to generate customer preferences). Uber also uses a non-relational database to store customer and driver data, trip history, trip details, and real-time location. Airbnb also uses No-SQL databases to store and manage data about properties available online, users, and a customer's accommodation history. This helps process very large amounts of data for a more complex and accurate analysis, thus helping improve the services offered on the site.

In order to illustrate more theoretical notions related to a No-SQL database, I will use MongoDB, as it is the most popular database (First place in the No-SQL database category and fifth place globally according to that ranking)[7].

Document-oriented databases have the *document* as their basic form. The document types supported by this database are: JSON (Java Script Object Notation), BSON (Binary Java Script Object Notation), XML (eXtensible Markup Language), YAML (Yet Another Markup Language). The structure of the documents is flexible, thus allowing a greater customization of the document content. Essentially, this format is an improved version of the No-SQL key-value database model, thanks to the fact that the document has a well-defined form. Redundancy in this case is allowed and even encouraged, since, in a No-SQL database, data is stored on multiple servers simultaneously. One advantage is the fact that it provides greater flexibility in searching and retrieving data, if a server in the network cannot be accessed, due to internal or external issues. This means less disruption to data access, which is essential and necessary for today's businesses. This principle is called *scalability*.

Another important aspect, nowadays, is the speed of data writing and reading. The everyday person wants to have fast access, preferably instant, to as much data as possible simultaneously, for various fields of activity (analysis, processing). This is provided by the lack of a database diagram, the lack of normalization forms, which allows the fast development of applications with No-SQL databases. Within document databases (MongoDB), the concept of junction does not exist. An equivalent of the junction concept is nesting an action through which a new document is inserted within the reference document. This significantly helps the processing of the document. This principle is called *efficiency*.

The variety of existing No-SQL database types has not achieved a standardisation of languages. In fact, this is an advantage, as it allows programmer communities and software development companies to compete with each other, to generate creative and unique ideas, ideas that can translate into new and innovative solutions, thus bringing progress and innovation in the field of No-SQL databases. Although the field of non-relational databases is quite new, compared to that of relational databases (which started in the late 1960s, more precisely in 1969[9]), this domain has known an impressive development. Today, many companies use a non-relational database as support, as we have seen above. This feature of non-relational databases is called *open-source*.

Non-relational databases are suitable for those who want to discover the universe of database knowledge and implicitly that of data analysis. The theoretical concepts related to No-SQL databases are fewer and much easier to understand. No-SQL database terminologies are easier to grasp, thus allowing for faster understanding and in-depth learning.

**Table 1.** Comparison of concepts between relational and non-relational databases

	<b>Non-relational (No-SQL)</b>	<b>Relational (SQL)</b>
Data storage	In tables populated with data - built based on the relational model	Storage using non-relational formats
Junctions	Yes	No
Data storage	Predefined types are respected	There are no defined forms, so it is semi-structured or unstructured.
Primary key	Yes	Yes, to which ObjectID or DocumentID can be added (MongoDB)
Diagram	Dynamic, adaptable depending on the reference model	Static
Transaction management	ACID Model Atomicity Consistency Isolation Durability	CAP Model Consistency Availability Partition Tolerance
Entity name	Table	Collection

BigData is a very important field, developed over the past years. Nowadays, this field is hugely useful, because the analyses and results provided by processing huge volumes of data help improve business processes, optimize decision-making processes, help with economic and financial forecasts, etc. Thus, the BigData field designates extremely large and complex data sets, which cannot be managed or analysed efficiently using traditional data processing tools, such as spreadsheets.[10]

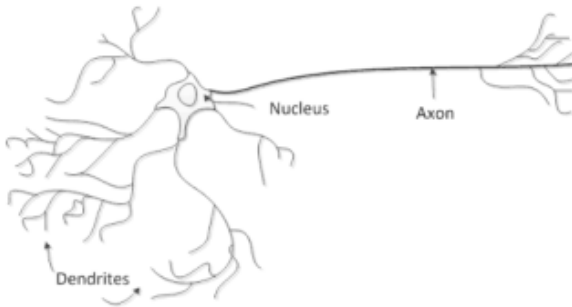
The increasing access to the Internet, economic growth and the increase of living standards worldwide, have irreversibly shaped the course of humanity. Through all his interactions (social, professional, economic, etc.) the Man becomes a data provider. Between 2012 and 2013, it was estimated, according to the IBM corporation, that 2.6 exabytes of data[11] were being produced daily, which means ( $2.17 \times 260$  bytes). Given this huge amount of data, in order to be able to process it, there are some steps that any analyst must take in order to reach conclusive and relevant results: data collection, data processing, data systematisation and then their analysis.

Data collection involves searching for data

in different sources of interest. The resulting data can be structured or unstructured. Data processing is the second stage in the set of operations that the collected data goes through. This operation involves the transformation and processing the data into a standardised and easy-to-process format. The next stage is data systematization. This process involves the removal of duplicates from the data set, the removal of outliers, which, after processing, are lost, they become corrupted, inaccessible, etc. This is an essential stage for maintaining the unity and consistency of the data, so that the result provided is as relevant and appropriate as possible. The last stage, data analysis, involves the application of one or more algorithms. Some of the algorithms used for analysis are: linear regression, logistic regression, k-clustering, k-nearest neighbours.[12] In order see how these algorithms are applied and used, the implementations will be shown place in the next chapter.

Besides, the field of BigData is the basis of AI (Artificial Intelligence). Artificial intelligence is the science aiming to make machines capable of performing tasks that would require intelligence, if they were performed by humans.[13],[17]. This

definition was provided by Marvin Minsky in 1956, during a scientific conference at Dartmouth College. BigData provides the information support on which this concept is developed. This is a field that has developed over the recent years. Four models of artificial intelligence have been discovered so far: Machine Learning (ML) which is based on the principle of autonomous learning. This is done based on the introduced data set, regardless of the field (mathematical, computer science, psychological, economic, etc.). Generative AI is the most advanced model known to date. Its characteristics are: generating images, videos, sounds or code written in different programming languages. These functionalities come as a result of a more focused and deeper learning, but also as a result of integrating more advanced languages (LLM). The other two models are: neural networks and deep learning (DL). Neural networks are reconstructions of the human brain at the algorithm level. Through this concept, an attempt is made to detail and deepen the functions of the human brain. Therefore, the starting point of these notions is represented by theoretical biology concepts. The human brain is made up of multiple neurons. A neuron is a cell specialised in receiving and transmitting information [14], [18]. Structurally, a neuron comprises: dendrites, a nucleus, and an axon. Synapses are the points of contact between two neurons, where the transmission of nerve impulses occurs. The following figure shows the structure of a neuron within the human brain:



**Fig. 1.** The structure of a neuron in the human brain

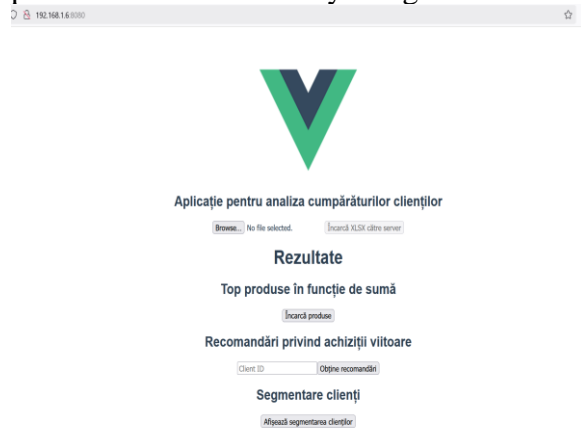
Starting from these concepts, the first abstraction of a neural network has been created: the synapses became *inputs*, the axon became *an output*, and the nucleus became a *summator*.

Thus, the BigData field has been in a permanent and continuous evolution. Most likely, in the future we will witness even more theoretical notions and concepts that are unknown today.

#### 4 NoSQL databases in the fields of Big Data and AI

In this chapter, we will present a case study: building a web application supported by a No-SQL database, provided by MongoDB, which will retrieve data about some customers from an XLSX file, and, data will then be recorded in the database. Based on the analyses made, the web programme will determine the customer's behaviour, thus suggesting offers for future purchases. This application can be useful for the marketing department of an e-commerce store. In order create the user interface, we will use Vue.js, and, in order to create the server connecting the database to the interface, we will use Node.js based on the JavaScript language. The advantage of web applications lays in their flexibility, as they can be run on any type of device (tablet, mobile phone, laptop, desktop, etc.), but also because most of these devices can be connected to the internet. Besides, another advantage of web applications resides in the variety of programming languages or frameworks available on the market to satisfy the needs of each programmer. Among the most famous frameworks we can mention: Vue.js, React.js, ASP.NET, Angular.js, Express.js, etc. The Vue.js framework will allow us to build a modern and user-friendly web interface. Vue.js is based on the HTML, CSS and JavaScript languages, for building objects (buttons, text writing areas, etc.) and for defining their behaviours (performing the operation shown by the button, etc.). CSS has the role of styling the web page with a modern,

pleasant and user-friendly design.



**Fig. 2.** Example of a user interface made with Vue.js

All of the application's operations are performed by the server (reading the XLSX file, inserting data into the MongoDB collection and the analytical data processing). In order to exemplify one of the analytical operations applied to the data set, such as ranking the most frequently purchased products, we will use, as a basis the following code fragment made in Node.js:

```
app.get("/analytics/top-products", async (req, res) => {
  let client;
  try {
    client = await connectToDb();
    const db = client.db(dbName);
    const collection = db.collection(collectionName);

    const topProducts = await collection
      .aggregate([
        { $group: { _id: "$ProductID", totalSales: { $sum: "$PurchaseAmount" } } },
        { $sort: { totalSales: -1 } },
        { $limit: 5 },
      ])
      .toArray();

    res.json({ topProducts });
  } catch (error) {
    console.error("Eroare la preluarea produselor:", error);
    res.status(500).send({ error: "Eroare de server" });
  } finally {
    if (client) await client.close();
  }
});
```

**Fig. 3.** Route making a connection to MongoDB, then applying processing operations, using MongoDB aggregations.

This route connects to the database, then to the desired collection. An aggregation is then applied using the `.aggregate(...)` operator. Then, using the `$group` operator, the records are grouped according to the `$ProductID` field. After that, the amount

spent by the buyer during the analysed period is calculated. The `$sort` operator sorts the data set in descending order, according to totalSales. Subsequently, the data is sent to the frontend, in JSON format, where it will be displayed. The frontend implementation will have two components: a JavaScript part represented by a function that will retrieve the data and a HTML part that will manage the display on the web page. The function retrieving the data from the server is the following:

```
async fetchTopProducts() {
  try {
    const response = await
    axios.get("http://localhost:3000/analytics/top-products");
    this.topProducts =
    response.data.topProducts;
  } catch (error) {
    console.error("Error fetching
    top products:", error);
  }
}
```

The specific HTML code for providing the final result is:

```
<section>
  <h2>Top produse în funcție de
  sumă</h2>
  <button
  @click="fetchTopProducts">Încarcă
  produse</button>
  <ul v-if="topProducts.length">
    <li v-for="product in
    topProducts" :key="product._id">
      ID: {{ product._id }} -
      Total vânzări: {{ product.totalSales
    }}
    </li>
  </ul>
</section>
```

In the result retrieval part, we used the Axios dependency. This dependency is useful when we want to write, delete, overwrite or obtain a data set. In this case, we want to retrieve the data using the `axios.get()` method. Then, we handle the possible errors that may occur during data retrieval. In the area where we display the result, by clicking on the *Incarca produse* (*Load products*) button, we will be able to see the results, because the button is marked with the `@click="fetchTopProducts"` event which



will refer to the JavaScript function. After that, a list of the elements obtained from the data retrieval, will be drawn up. The result will have the following form

### Top produse în funcție de sumă

Încarcă produse

ID: P1003 - Total vânzări: 599.99  
 ID: P1009 - Total vânzări: 499.99  
 ID: P1004 - Total vânzări: 449.98  
 ID: P1007 - Total vânzări: 399.99  
 ID: P10012 - Total vânzări: 399.99

**Fig. 4.** Example of result within the web application

The *k-nearest-neighbours* algorithm is one of the most well-known algorithms in the field of BigData processing and in the field of artificial intelligence. This algorithm is used in classification and regression problems. It is considered one of the easiest algorithms in the field of *machine learning*. The implementation stages of this algorithm are: choosing the number of *k* neighbours closest to a reference point. Then, the distance between any two points in the plane is calculated until the entire range of points is covered. The most commonly used formula to calculate the distance is the Euclidean distance. There are other types of distances as well: Manhattan distance, Minkowski distance, Hamming distance: which is used for categorical data. After calculating the distance, the shortest *Ks* to the reference point are chosen.

Thus, we will implement this algorithm to highlight the following scenario: generating the potential preferences of a customer in the store based on their previous purchases. Let's assume we choose *k=5*, that is, the number of suggestions for the customer. Therefore, we will build the function calculating the distance, with the formula for the Euclidean distance.

```
const calculateDistance = (vectorA,
vectorB) => {
```

```
    return
    Math.sqrt(vectorA.reduce((sum, val, i)
=> sum + Math.pow(val - vectorB[i],
2), 0));
};
```

Now, we implement the route within the server. Due to the fact that the code is very large, I will present it in several different figures, as follows:

```
app.get("/analytics/knn-recommendations/:clientId", async (req, res) => {
  const clientId = parseInt(req.params.clientId);
  const k = parseInt(req.query.k) || 5; // Setăm implicit numărul de sugestii implicit
  la 5
  let client;

  try {
    client = await connectToDb();
    const db = client.db(dbName);
    const collection = db.collection(collectionName);

    // Preluăm toți clienții cu cumpărăturile lor

    const clients = await collection.find().toArray();
    const targetClient = clients.find((c) => c.ClientID === clientId);

    if (!targetClient) {
      return res.status(404).send({ message: "Clientul nu s-a găsit." });
    }

    // Construim vectorul caracteristic unde vom salva sugestiile

    const featureVectors = clients.map((c) => ({
      clientId: c.ClientID,
      vector: [
        c.PurchaseAmount || 0, ... (c.ProductCategory ?
[c.ProductCategory.length] : [0]), ],
    }));

    const targetVector = [
      targetClient.PurchaseAmount || 0,
      ... (targetClient.ProductCategory ? [targetClient.ProductCategory.length] : [0]),
    ];
```

**Fig. 5.** Part I – connecting to the database and building the characteristic vector.

```
// Calculăm distanțele folosindu-ne de funcția specifică a calcului distanței

const distances = featureVectors
  .filter((fv) => fv.clientId !== clientId)
  .map((fv) => ({
    clientId: fv.clientId,
    distance: calculateDistance(targetVector, fv.vector),
  }));

// Găsim cei mai apropiați K vecini

const nearestNeighbors = distances
  .sort((a, b) => a.distance - b.distance)
  .slice(0, k);

// Preluăm datele celorlalte clienți

const recommendations = [];
for (const neighbor of nearestNeighbors) {
  const neighborPurchases = await collection.find({ ClientID: neighbor.clientId
}).toArray();
  recommendations.push(...neighborPurchases.map((p) => p.ProductID));
}

res.json({ recommendations: [...new Set(recommendations)].slice(0, 5) }); //
Returnăm noile sugestii

} catch (error) {
  console.error("Eroare:", error);
  res.status(500).send({ error: "Eroare de server" });
} finally {
  if (client) await client.close();
}
});
```

**Fig. 6.** Calculating the distance and selecting the neighbours.

Now, we will see the result for a customer,

having an ID with the value of 1001.

```

1  |
2  | "recommendations": [
3  |   "P2001",
4  |   "P5003",
5  |   "P9003",
6  |   "P5007",
7  |   "P10010"
8  | ]
9  |

```

**Fig. 7.** Example of result within the web application using the k nearest neighbours algorithm

Of course, these algorithms can also be used in other fields of study (medicine, education, networking, etc.), or, in our case, for other scenarios: optimising results based on preferences/Internet searches/budget allocated by the consumer. This way, the online advertising provided by merchants will be more relevant to the consumer. In order to have the expected, desired results, we must have as much data as possible, which also has to be relevant. The premises that we want to analyse must be logical and they must have a well-defined purpose, so that the decision-maker can always make the best decision.

## 5 Conclusions

In conclusion, No-SQL databases have greatly contributed to deepening and consolidating the BigData field. With their flexible and scalable way of managing data, non-relational databases prove that they can be an alternative to traditional relational databases, which have a more rigid structure and which are not so flexible, when it comes to processing large volumes of data. MongoDB databases also stand out in BigData processing, as they provide a new perspective on the process of storing and analysing data. The use of data collections and nesting within documents allows for easy and fast data management, compared to the traditional relational models, which do not benefit from so much flexibility and contain more

theoretical concepts, as seen in the previous chapters. Besides, the use of operators in data processing is an advantage, as they are intuitive and easy to use, being of great help to new programmers in this field. In addition, the study of No-SQL databases is an important step for any programmer, entering a new path of knowledge. Besides, the syntax of the MongoDB operators, is also easy and intuitive. Machine-learning algorithms along with the processing of large volumes of data make up a symbiosis that will develop the study and research in any scientific field, providing as many solutions as possible, to questions that have not yet been answered.

## References

- [1] <https://www.telacad.ro/cursuri/baze-de-date-nosql-redis-mongodb-neo4j/>
- [2] <https://www.oracle.com/ro/database/nosql/what-is-nosql/>, accessed on 21.03.2025
- [3] <https://www.guru99.com/ro/sql-vs-nosql.html>, accessed on 21.03.2025
- [4] <https://ro.simeononsecurity.com/article/s/sql-vs-nosql-choosing-the-right-database-management-system/>, accessed on 21.03.2025
- [5] <https://www.unite.ai/ro/ce-este-un-inginer-de-date-foaie-de-parcurs-responsabilit%C4%83%C8%9Bi-salariale/>, accessed on 21.03.2025
- [6] coord. I. Lungu, Adela Bâră, Constanța Bodea, Iuliana Botha, Vlad Diaconița, Alexandra Florea și Anda Velicanu, *Tratat de baze de date vol. I: Baze de date, organizare, proiectare, implementare*, Editura ASE, București, 2011, p. 77
- [7] <https://db-engines.com/en/ranking>, accessed on 12.12.2024
- [8] <https://www.datastax.com/guides/nosql-use-cases>, accessed on 12.12.2024
- [9] coord. I. Lungu, Adela Bâră, Constanța Bodea, Iuliana Botha, Vlad Diaconița, Alexandra Florea și Anda Velicanu, *Tratat de baze de date vol. I: Baze de date, organizare, proiectare*,

- implementare*, Editura ASE, București, 2011, p. 77
- [10] <https://www.oracle.com/ro/big-data/what-is-big-data/>, accessed on 15.12.2024
- [11] <https://web.archive.org/web/20130824213031/http://www.ibm.com/big-data/us/en/>, accessed on 16.12.2024
- [12] <https://tdwi.org/articles/2018/07/02/adv-all-5-algorithms-for-big-data.aspx>, accessed on 17.12.2024
- [13] <https://www.sap.com/romania/products/artificial-intelligence/what-is-artificial-intelligence.html>, accessed on 17.12.2024
- [14] <https://code-it.ro/introductere-in-retele-neuronale-teorie-si-aplicatii/>, accessed on 18.12.2024
- [15] <https://codegym.cc/ro/quests/lectures/ro.quests.hibernate.level19.lecture00>, accessed on 21.03.2025
- [16] <https://support.microsoft.com/ro-ro/topic/no%C8%9Biuni-elementare-despre-bazele-de-date-a849ac16-07c7-4a31-9948-3c8c94a7c204>, accessed on 21.03.2025
- [17] <https://www.europarl.europa.eu/topics/ro/article/20200827STO85804/ce-este-inteligenta-artificiala-si-cum-este-utilizata>, accessed on 21.03.2025
- [18] <https://askabiologist.asu.edu/neuronii-%C5%9Fi-nervii>, accessed on 21.03.2025



**CHIRIAC Mario-Tudor** First year student, in the Master's Programme named: "Databases - business support", within the Faculty of Cybernetics, Statistics and Economic Informatics - the Bucharest University of Economic Studies. In 2024, I graduated from the Bachelor's programme - Economic Informatics in Romanian, within the Faculty of Cybernetics, Statistics and Economic Informatics at the Bucharest University of Economic Studies. Areas of interest: study of SQL and No-SQL databases, programming web applications using Vue.js, React.js.

# House Market Prediction Using Machine Learning

Nicușor-Andrei ANDREI

Bucharest University of Economic Studies

Faculty of Cybernetics, Statistics and Economic Informatics

Bucharest, Romania

[andreinicusor20@stud.ase.ro](mailto:andreinicusor20@stud.ase.ro)

*This study explores and compares the performance of tree-based machine learning algorithms' predictions for Bucharest real estate market prices. The dataset was obtained from a local platform in March 2025 and contains residential apartments for sale in Bucharest. The comprehensive data preprocessing step, including imputation of missing values, encoding of categorical variables, and the engineering of new key features such as distance to public transport, played a key role in the models' performance. The models were optimized using a grid search algorithm with 5-fold cross-validation and evaluated with key performance indicators including root mean squared error, mean absolute error, and coefficient of determination. The results indicate that XGBoost outperforms both Random Forest and a single Decision Tree, reducing all the key performance indicators used in analysis.*

**Keywords:** House market, Machine learning, Price prediction, Tree-based machine learning, XGBoost

## 1 Introduction

The housing real estate market in Bucharest underwent a drastic transformation over time due to rapid urbanization, demographic changes, and shifting economic conditions. Homebuyers and investors require an effective price forecast based on newer techniques to provide a robust house market analysis. The performance of tree-based machine learning techniques, decision tree, random forest, and XGBoost, on this particular economic segment of Bucharest is presented in this paper. One Decision Tree provides simple interpretability but will probably overfit; Random Forests increase stability by averaging a large number of trees; and XGBoost continues to improve predictions using gradient boosting that adjusts residual errors iteratively. Despite the optimistic results presented by these approaches in different global markets, their comparative performance in the specific context of Bucharest is yet to be explored. The current research fills this gap by benchmarking the performance of Decision Tree, Random Forest, and

XGBoost regressors against an elaborate dataset of real estate transactions in Bucharest, collected in March 2025 from a regional online marketplace. After an extensive preprocessing step including imputation of the missing values, outliers handling, and feature engineering, the algorithms are fine-tuned using the grid search method. The grid search is carried out using a 5-fold cross-validation mechanism with the best model determined as the minimizer of the mean squared error metric. The optimised versions of the three algorithms are next benchmarked against the key indicators relevant to regression analysis: root mean squared error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), and the coefficient of determination ( $R^2$ ). Apart from precision, the approach also considers the drivers of the prices of properties, with model interpretability highlighted using the feature importance score.

The structure of the paper is outlined as follows: Section 2 gives an overview of the literature relevant to the topic under investigation; Section 3 gives a comprehensive description of the

methodology used during this research; Section 4 presents the findings together with a related discussion; and Section 5 recapitulates the major conclusions of this research and makes recommendations for future research. This framework ensures an empirically supported comparison of tree-based models in the context of Bucharest's vibrant real estate market.

## 2 Literature Review

The rapid evolution of machine learning has opened new opportunities for modeling complex systems, including the real estate market, where price dynamics are difficult to capture. In the context of Bucharest's urban housing landscape, machine-learning-based price prediction promises to help buyers, investors, and policymakers navigate a market shaped by diverse offers and fluctuating demand. In the next paragraphs, studies related to this field are synthesized to highlight methodologies, key findings, and their implications for a Bucharest-focused investigation.

### 2.1 Hedonic Foundations

Before the beginning of modern machine learning, hedonic-price models provided a framework for decomposing sale prices into intrinsic and external attributes [1]. Early applications in Beijing and Paris confirmed the influence of building quality and neighborhood heterogeneity on prices [2, 3], and more recent work has extended hedonic indices to capture the effects of transit accessibility [4]. Although most hedonic studies focus on explanatory power rather than forecasting, they offer valuable guidance on feature selection and baseline model structure, insights that remain relevant when constructing machine-learning pipelines for Bucharest.

### 2.2 Ensemble and Tree-Based Methods

A recurring theme across multiple markets is the strong performance of tree-based ensemble models. For example, the authors of [5, 6, 7] show that Random Forest Regression achieves high accuracy when

forecasting house prices in their case studies. Similarly, in Hong Kong, a Random Forest model outperformed both Support Vector Machines and Gradient Boosting Machines on a dataset of 40,000 transactions spanning 18 years [8]. Extreme Gradient Boosting algorithm (XGBoost) also stands out: a Vilnius study that scraped nearly 19,000 listings found XGBoost to be the most predictive among fifteen tested models [9]. In Bengaluru, XGBoost again outperformed Linear, Ridge, Lasso and SVM baselines [10], and hybrid approaches (for example stacking with CatBoost) further reduced error on 1.9 million transactions in Hong Kong [11]. Taken together, these ensemble successes suggest that machine-learning models for Bucharest will likely benefit from tree-based learners, provided that sufficient feature engineering and hyperparameter tuning are applied.

### 2.3 Linear and Kernel Methods

Although less dominant than ensemble models, linear and kernel-based regressors remain valuable for their interpretability and computational efficiency. In the Bengaluru comparison, Ridge and Lasso regressions, as well as a support-vector machine, delivered reasonable accuracy but were ultimately outperformed by XGBoost [10]. In Melbourne, an SVM achieved the lowest mean squared error at the expense of the longest training time, underscoring a trade-off between predictive power and computational cost [12]. A Multiple linear regression with the right data split ratio is validated as being more efficient than simple linear regression for house market predictions [13]. In Beijing, after a rigorous data preprocessing step, which included outlier elimination, feature engineering, and one-hot encoding algorithm, Hybrid and Stacked Generalization Regression delivered promising results on the training set and test set [14].

For Bucharest, where data volumes may vary significantly across districts, it may be

worthwhile to benchmark a well-tuned SVM alongside faster linear models as baseline approaches.

## 2.4 Classification Formulations

Some researchers have reframed house-price forecasting as a binary classification task, predicting whether a property's selling price will exceed its listing price. In Fairfax County, Virginia, the RIPPER rule-learning algorithm yielded fewer classification errors than C4.5, Bayesian classifiers, and AdaBoost [15]. A related study applied SVM, Random Forest, and a neural network, finding that Random Forest delivered the highest accuracy, precision, sensitivity, and specificity [16]. While classification methods do not provide information on the magnitude of price changes, they may prove useful in Bucharest for stakeholders interested primarily in upside versus downside risk.

## 2.5 Deep and Hybrid Architectures

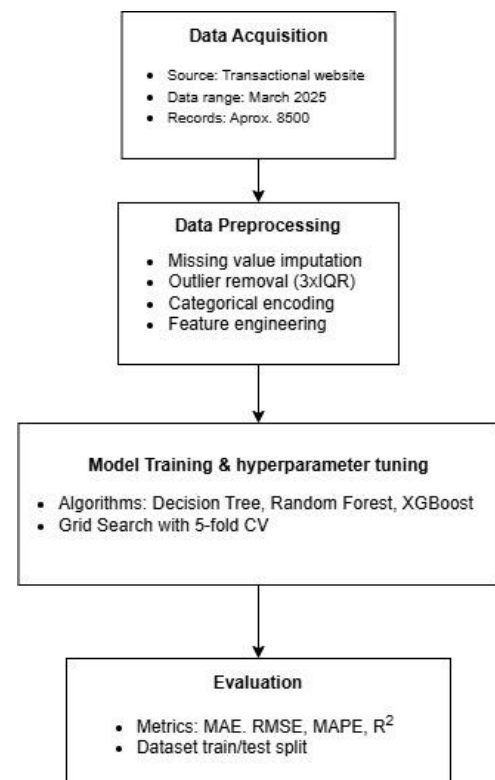
Deep-learning models, especially when enriched with a good data engineering step, have great potential but also require large datasets and careful regularization to avoid overfitting. One study combined street-view imagery and socio-economic indicators with a gradient boosting machine to forecast neighborhood-scale appreciation rates with high accuracy [17]. Comparative work found that both deep networks and classical regression models tended to overestimate control-sample prices, indicating a need for larger datasets and ensemble hybrids [18]. The most advanced approaches leverage Transformer architectures optimized via Bayesian hyperparameter search, achieving substantial RMSE reductions on large Hong Kong datasets [11]. For Bucharest, integrating aerial or street-view data, if such data are available, could yield similar improvements, although attention must be paid to model complexity and generalization.

## 2.6 Research Gaps and Outlook

Despite the extensive literature, few studies have applied these advanced machine-learning techniques explicitly to the Bucharest housing market. Key knowledge gaps include the integration of geographic data with house attributes and the adaptation of machine learning models and deep-learning networks to conform to the data structure of Romania. By overcoming these limitations and integrating best practices across the different stages of preprocessing of the data, model development, and model evaluation, future studies can provide accurate and interpretable price forecasts customized to the specific urban context of Bucharest.

## 3 Methodology

The following chapter, summarized in Figure 1, presents the methodology steps in depth, starting with data acquisition and preprocessing, continuing with the architectures of the proposed models, and closing with the metrics used for comparison.



**Fig. 1** Methodology applied



### 3.1 Data Acquisition and Preprocessing

Transactional data for the Romanian residential real estate market was obtained by web scraping a local listings website in March 2025. Initial preprocessing involved the removal of variables presenting over 5% missing values. Discrete features with missing entries were imputed according to domain-specific rules: absent bathroom counts were assigned a value of 1, and missing balcony counts were set to 0. The year built field, with a low percentage of missing values, was median-imputed to maintain discreteness, while missing “built area” values were replaced by the median built area stratified by room count.

Feature engineering produced several binary and continuous variables to enhance model performance. The categorical attributes `housing_type`, `comfort`, and `compartmentalization_type` were encoded as the binary indicators `is_penthouse`, `is_luxury`, and `is_detached`. Vertical position within the building was captured by the flags `is_first_floor` and `is_last_floor`, together with `floor_ratio`, defined as the apartment’s floor number divided by the total building floors.

Spatial features were derived from latitude and longitude coordinates by computing distances (in meters) to the city center and the nearest metro station using the Haversine formula [19]:

$$\varphi_i^* = \varphi_i \cdot \frac{\pi}{180}, \quad \lambda_i^* = \lambda_i \cdot \frac{\pi}{180}, \quad i = 1, 2 \quad (1)$$

$$\Delta\varphi = \varphi_2^* - \varphi_1^*, \quad \Delta\lambda = \lambda_2^* - \lambda_1^* \quad (2)$$

$$a = \sin^2\left(\frac{\Delta\varphi}{2}\right) +$$

$$\cos(\varphi_1^*) \cos(\varphi_2^*) \sin^2\left(\frac{\Delta\lambda}{2}\right) \quad (3)$$

$$d = 2 R \arcsin(\sqrt{a}), \quad (4)$$

where  $\varphi$  represents latitude,  $\lambda$  represent longitude and  $R$  represents the radius of the Earth, which is 6371 km. [20]

Interquartile filtering was applied to mitigate the influence of extreme outliers. For each variable of interest—sale price, usable area, built area, bathroom count, and floor count—the first (Q1) and third (Q3) quartiles were calculated, and the interquartile range ( $IQR = Q3 - Q1$ ) was

determined. Observations falling outside the interval  $[Q1 - 3 \cdot IQR, Q3 + 3 \cdot IQR]$  were removed, resulting in a homogeneous and robust dataset suitable for subsequent inferential and predictive modeling.

### 3.2 Model Architectures

This study presents a comparative analysis between the Decision Tree, Random Forest, and XGBoost regression algorithms. To minimize the mean squared validation error, the exhaustive grid search approach was employed to select the optimal hyperparameters for each model, with the aid of 5-fold cross-validation. The exhaustive grid search provided the best parameter settings that achieved the lowest average validation MSE for each of the respective algorithms, thus attaining a good trade-off between bias and variance [21].

The Decision Tree algorithm partitions the feature space by using splits that increase variance reduction according to the squared-error criterion. To avoid the over-splitting and overfitting, three complexity parameters were tuned as explained below:

- **max\_depth:** (3, 5, 7, 10, 12, 15, 20, 25) - Limits the depth of the tree to balance representation ability with overfitting risk.
- **min\_samples\_split:** (2, 5, 10, 20) - Specifies the minimum number of samples to be used when making a split for a node.
- **min\_samples\_leaf:** (1, 2, 4, 10) - Guarantees terminal nodes to have a reasonable quantity of observations, smoothening predictions in regions of sparsity.

The Random Forest algorithm builds a series of decision trees by bootstrap aggregation (bagging) and random feature selection, thus reducing variance by averaging [22]. Four hyperparameters were explored:

- **n\_estimators:** (10, 50, 100, 200, 400, 600, 800) - Controls the size of the forest of trees; larger forests



reduce variance more effectively at increased computational cost.

- **max\_depth:** (3, 5, 7, 10, 12) - Max out the depth of each tree to control individual complexity.
- **max\_features:** ("sqrt", "log2") - Controls the number of predictors considered at each split, introducing randomness which decorrelates individual trees further.
- **bootstrap:** (True, False) - Switches sampling with or without replacement, evaluating the effect of bootstrap aggregation on ensemble stability.

This extensive search guaranteed the choice of a forest configuration that minimized validation error while ensuring computational efficiency. XGBoost uses gradient tree boosting to train trees sequentially on the residuals of earlier iterations, with a squared-error loss combined with regularization to avoid overfitting. [23] Three of the most important parameters were optimized:

- **learning\_rate:** (0.01, 0.1, 0.3) - Scales the contribution of each new tree, with lower values promoting slow learning and improved generalization.
- **max\_depth:** (3, 5, 7, 10, 12) - Cuts off the depth of each boosted tree, balancing the ability to model complex patterns against overfitting.
- **n\_estimators:** (10, 50, 100, 200, 400, 600, 800) – The total number of trees to be trained, more iterations allowing for more accurate residual correction, but also using more training time.

By applying the grid search technique over these grids, the optimal balance among learning rate, tree complexity, and ensemble size was determined.

### 3.3 Evaluation Measures

Models' performance was compared on an unseen test set of data against four

complementary metrics chosen for their individual interpretive strengths in housing-market forecasting contexts: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and the coefficient of determination ( $R^2$ ).

Root Mean Squared Error was selected since it punishes huge deviations more severely than small ones, aligning with the fact that huge pricing errors (e.g., underpricing a home by tens of thousands euros) are particularly unperformant for real-estate applications. By squaring residuals before averaging, RMSE rewards models that avoid occasional but big mispredictions, so it is a sensitive metric of worst-case performance.

Mean Absolute Error provides a straightforward, linear measure of average prediction error in the same currency units as the target, which is euros. Unlike RMSE, MAE treats all errors equally and is thus more robust to outliers and more interpretable for stakeholders who require an intuitive sense of actual deviation from the actual sale price.

To measure errors concerning scale, the Mean Absolute Percentage Error (MAPE) was applied. MAPE is a measure of the average deviation from the true price, thus allowing effective comparisons between segments with different levels of prices (e.g., comparing central business district penthouses with suburban apartments). The ratio measure is useful in providing decision-makers with the ability to measure the relative accuracy of the models.

Finally, the coefficient of determination quantifies the degree to which observed price variation is accounted for by the model. A value of  $R^2$  close to 1 signifies strong explanatory power for market movements, while values close to zero imply that the predictor set fails to adequately capture systematic pricing patterns. Together, these four measures offer a complete and balanced evaluation of accuracy, robustness, relative error, and

goodness-of-fit, and thus allow for a strict comparison of the Decision Tree, Random Forest, and XGBoost models.

## 4 Results and Discussions

This chapter presents an in-depth view of both dataset features and exploratory analysis, along with each individual model's performance, and concludes with a comparative view of the trained models.

### 4.1 Dataset Presentation

The final dataset used in the analysis has 8500 entities, each with the following independent variables based on type:

- Continuous variables:
  - Built\_area
  - Usable\_surface
  - Lat
  - Lon
  - Metro\_distance\_in\_meters
  - Distance\_to\_center
  - Floor\_ratio
- Integer Variables:
  - Year\_built
  - Floor\_number
  - Number\_of\_floors
  - Bedroom\_count
  - Bathroom\_count
- Dummy variables:
  - Is\_penthouse
  - Is\_luxury
  - Is\_first\_floor
  - Is\_last\_floor
  - Is\_detached

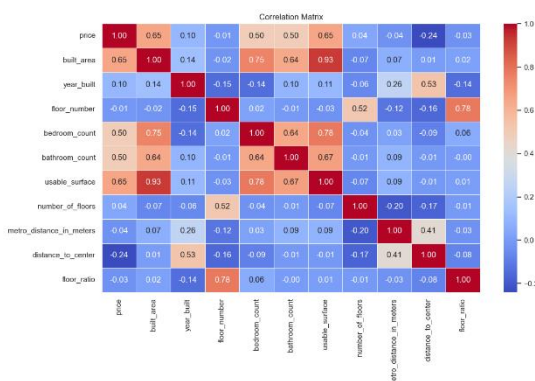


Fig. 2. Correlation Matrix

The correlation table in Figure 2 gives a general overview of the association

between predictors and between predictors and the target variable. There is a moderate positive relationship between price and bathroom\_count, bedroom\_count, and usable\_surface, which reflects the major role of building size and layout in influencing the price. In contrast, there is a weak negative relationship between distance\_to\_center and price, which shows that an increase in distance from the city center has the general tendency to lower the value of the properties. Furthermore, distance\_to\_center and metro\_distance\_in\_meters have moderate positive relationships with year\_built, which shows that newer constructions are most likely to be located far from the city center and metro stations. Finally, the strong relationship between usable\_surface, built\_area, bedroom\_count, and bathroom\_count supports the fact that larger living spaces usually have more bedrooms and bathroom counts. These findings reinforce the significant part played by structural and location variables in establishing the value of Bucharest residential properties.

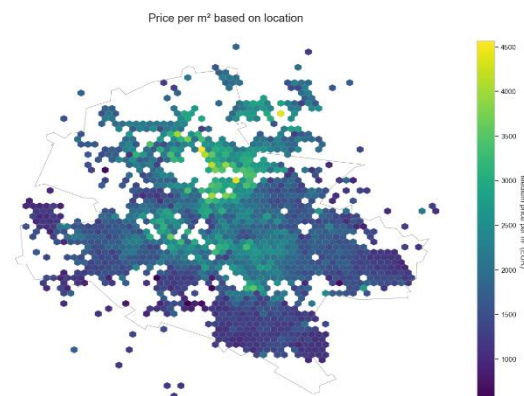


Fig. 3. Price per m2 based on the location of the apartment

The hexbin map in Figure 3 illustrates price per square meter variability across the geography of the city. The higher values cluster together in the north-central suburbs, where there is high access speed to central amenities and a well-connected transport network, increasing demand. Prices fall off gradually from the center,

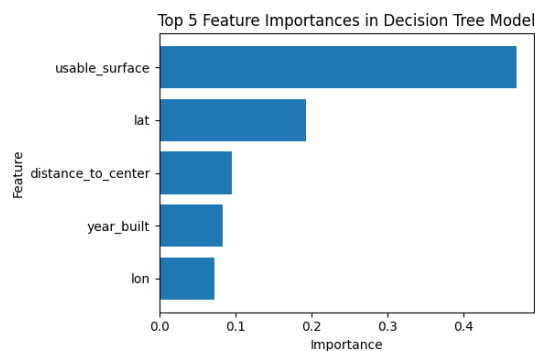
then fall lowest in the southern, eastern, and western peripheries—suburbs dominated by newly developed areas that haven't yet had time to experience the benefits of a fully settled urban infrastructure.

#### 4.2 Models Overview

This chapter presents the final resulted models by the grid search algorithm along with their parameters and feature importance.

For the Decision Tree Regressor, the optimal model based on the MSE value resulting from the grid search algorithm has the following parameters:

- Max\_depth: 10
- Min\_samples\_leaf: 10
- Min\_samples\_split: 2



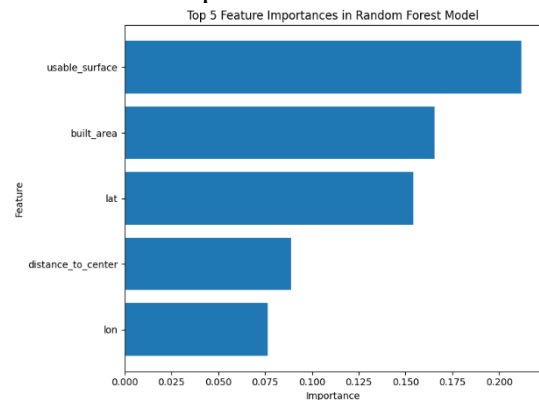
**Fig. 4.** Top 5 feature importances in Decision Tree Regressor

The barplot presented in Figure 4 presents the top 5 characteristics by importance, which are used by the decision tree model to predict the final price of the residential building. The most important variable used in forecasting is the usable surface, but the geospatial data represented by latitude and longitude, along with the distance from the city center, also play a key role in predictions.

Based on the same search algorithm, the optimised Random Forest algorithm has the following parameters:

- Max\_depth: 12
- Max\_features: 'sqrt'
- N\_estimators: 200

- Bootstrap: False

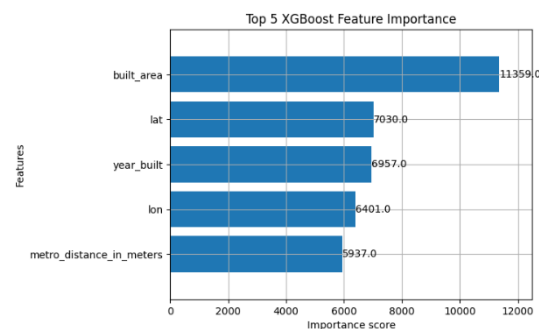


**Fig. 5.** Top 5 feature importances in Random Forest Regressor

Figure 5 illustrates the top 5 features by importance used by the Random Forest model. The barplot also suggests that distance to city center, geospatial location, and the surface of the apartment are among the most important features used for price prediction. Regarding the Decision tree model, the Random forest model uses the built area of the apartment instead of the year built as one of the top 5 features.

For the XGBoost ensemble model, the optimal parameters after training using the grid search algorithm are:

- Learning\_rate: 0.1
- Max\_depth: 7
- N\_estimators: 800
- 



**Fig. 6.** Top 5 XGBoost feature importance

Figure 6 presents the top 5 features used by the XGBoost model in predictions. In the same manner as the last models, the surface of the residential unit, the location, and the distance to the center are the main factors used in predictions.

### 4.3 Model Comparison

**Table 1.** Performance of the models based on key indicators

	MAE	RMSE	MAPE	R2
Decision Tree	20487.65	32589.67	15.86	0.77
XGBoost	15089.36	24296.89	11.88	0.87
Random Forest	17134.86	26573.65	13.7	0.85

This chapter will analyze and compare the performance of the studied models presented in Table 1. The single Decision Tree model, although very interpretable, performs the worst in terms of prediction among the three methods. It has a Mean Absolute Error (MAE) of approximately 20,488 and a Root Mean Squared Error (RMSE) of 32,590 on the test set, meaning its point-predictions tend to be short of actual values by a significant margin. Its MAPE of 15.9 % shows that, on average, predictions are different from actual prices by nearly one-sixth, and its  $R^2$  of 0.77 shows that only 77% of the variability in sale prices is explained. This result illustrates the model's tendency to overfit patterns in the training data and not to smooth out noise between different regions of the feature space.

By comparison, the Random Forest ensemble substantially reduces the bias and variance through the application of bootstrap aggregation. Its MAE is reduced to approximately 17,135 (a decrease of 16 % relative to the Decision Tree), and its RMSE is reduced to 26,574, while the MAPE is reduced to 13.7 %. Such reductions yield an  $R^2$  equal to 0.85, indicating that 85 % of residential prices' variance is accounted for by the model. The improved performance comes from averaging the predictions of 200 decorrelated trees, each with a maximum depth of 12, which stabilizes estimates and lowers over-fitting at the expense of less interpretability.

Finally, the XGBoost model achieves the highest overall accuracy by sequentially correcting the residuals of its predecessor. It gives an MAE of 15,089 (26 % lower than the Decision Tree), an RMSE of

24,297, and a MAPE of just 11.9 %, which corresponds to an  $R^2$  of 0.87. Practically, this means the gradient-boosted ensemble reduces average percentage error by nearly one-quarter compared to the tree baseline and explains 87 % of price variation. Though slightly more complex to train and to tune, XGBoost's ability to learn subtle feature interactions and to penalize overly complex trees makes it the go-to when minimizing prediction error is top priority.

### 5 Conclusion

In this research, a thorough comparison of tree-based regression algorithms Random Forest, Decision Tree, and XGBoost for predicting apartment prices in residential apartments in Bucharest has been performed. Using a true transactional dataset collected in March 2025, which was subject to detailed preprocessing, missing-value imputation, one-hot encoded, and creating new spatial indicators, each model was optimized using grid search combined with five-fold cross-validation to reduce the mean squared error (MSE) as the goal function. Among the independent variables involved, geospatial variables like proximity to the city center, longitude, latitude, distance to metro station, and unit size measurements (usable area and built area) were the most critical indicators of the price.

The results reveal a clear and noteworthy trend in the predictive accuracy. The single Decision Tree, while illustrating higher interpretability, incurred the largest testing errors, with the mean absolute percentage error being 15.9%. Application of the bootstrap averaging in the Random Forest approach reduced the mean error by 16%,

thereby attesting to the importance of variance reduction via ensemble methods. Finally, the gradient-boosted XGBoost model reported the lowest errors, demonstrating its ability to accurately make incremental corrections and discern complex, non-linear relationships.

Practically, these findings suggest that ensemble methods, particularly XGBoost, are significantly better than single trees when reducing pricing error is the goal. Random Forest is a good compromise if interpretability and computational cost need to be balanced against performance. In the meantime, the Decision Tree remains a good benchmark for rapid prototyping and stakeholder communication because it has interpretable decision rules.

Despite these advancements, some of these limitations remain to be explored in future research. First, our spatial attributes depend on fixed coordinates and do not account for dynamic urban factors such as traffic congestion or planned infrastructure development. Second, temporal dynamics (for instance, seasonality, macroeconomic data) were not explicitly modeled; employing time-series methods or adding lagged market indicators might enhance predictions further. Investigating these directions will contribute to real-time appraisal systems and support data-informed decision-making in Bucharest's transitioning real estate market.

## References

- [1] S. Rosen, "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition," *Journal of Political Economy*, vol. 82, no. 1, 1974.
- [2] Z. Yang, "An application of the hedonic price model with uncertain attribute - The case of the People's Republic of China," *Property Management*, vol. 19, no. 1, pp. 50-63, 2001.
- [3] R. P. M. & S. S. Maurer, "Hedonic price indices for the Paris housing market," *Allgemeines Statistisches Archiv*, vol. 88, pp. 303-326, 2004.
- [4] Y. C. N. X. R. Z. K. C. S. H. Linchuan Yang, "Place-varying impacts of urban rail transit on property prices in Shenzhen, China: Insights for value capture," *Sustainable Cities and Society*, vol. 58, p. 102140, 2020.
- [5] M. a. J. a. Y. S. a. P. R. a. N. P. a. Indervati, "House Price Prediction Using Machine Learning," in *1st International Conference on Advances in Computing, Communication and Networking (ICAC2N)*, Greater Noida, India, 2024.
- [6] Y. L. M.-H. L. a. S.-Y. H. Jeonghyeon Kim, "A Comparative Study of Machine Learning and Spatial," *Sustainability*, vol. 14, no. 15, p. 9056, 2022.
- [7] O. N. A. F. A. A. O. O. Y. F. A. G. O. Abigail Bola Adetunji, "House Price Prediction using Random Forest Machine Learning Technique," *Procedia Computer Science*, vol. 199, pp. 806-813, 2022.
- [8] B.-S. T. S. W. W. Winky K.O. Ho, "Predicting property prices with machine learning algorithms," *Journal of Property Research*, vol. 38, no. 1, pp. 48-70, 2021.
- [9] P. V. & S. A. Grybauskas, "Predictive analytics using Big Data for the real estate market during the COVID-19 pandemic," *Journal of Big Data*, vol. 8, p. 105, 2021.
- [10] R. G. a. N. S. N. J. Manasa, "Machine Learning based Predicting House Prices using Regression Techniques," in *2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)*, Bangalore, India, 2020.
- [11] Y. L. Z. W. M. Z. T. W. C. Choujun Zhan, "A hybrid machine learning framework for forecasting house price," *Expert Systems with Applications*, vol. 233, p. 120981, 2023.

- [12] T. D. Phan, "Housing Price Prediction Using Machine Learning Algorithms: The Case of Melbourne City, Australia," in *2018 International Conference on Machine Learning and Data Engineering (iCMLDE)*, Sydney, NSW, Australia, 2018.
- [13] A. a. A. H. a. S. R. Rai, "Predicting Housing Sale Prices Using Machine Learning with Various Data Split Ratios," *Data and Metadata*, vol. 3, 2024.
- [14] M. N. H. D. B. M. Quang Truong, "Housing Price Prediction via Improved Machine Learning Techniques," *Procedia Computer Science*, vol. 174, pp. 433-442, 2020.
- [15] J. K. B. Byeonghwa Park, "Using machine learning algorithms for housing price prediction:," *Expert Systems with Applications*, vol. 42, no. 6, pp. 2928-2934, 2015.
- [16] B. a. S. Dutta, "Predicting the housing price direction using machine learning techniques," in *2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI)*, Chennai, India, 2017.
- [17] F. Z. W. P. S. G. J. R. F. D. C. R. Yuhao Kang, "Understanding house price appreciation using multi-source big geo-data and machine learning," *Land Use Policy*, vol. 111, 2021.
- [18] Foryś, "Machine learning in house price analysis: regression models versus neural networks," *Procedia Computer Science*, vol. 207, pp. 435-445, 2022.
- [19] R. A. A. a. F. Darnis, "Use of Haversine Formula in Finding Distance Between Temporary Shelter and Waste End Processing Sites," *Journal of Physics: Conference Series*, vol. 1500, 2020.
- [20] J. & H. S. Carroll, "Using a video camera to measure the radius of the Earth," *Physics Education*, vol. 48, no. 6, p. 731., 2013.
- [21] P.-B. G. W. Iwan Syarif, "SVM Parameter Optimization Using Grid Search and Genetic Algorithm to Improve Classification Performance," *TELKOMNIKA*, vol. 14, no. 4, pp. 1502-1509, 2016.
- [22] Y. W. Y. Z. J. Liu, "New Machine Learning Algorithm: Random Forest," in *Information Computing and Applications. ICICA 2012.*, 2012.
- [23] G. Tianqi Chen, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKD International Conference on Knowledge Discovery and Data Mining*, 2016.



**Nicușor Andrei** earned his bachelor's degree in Economic Informatics in 2023 at the Academy of Economic Studies from Bucharest. He is a master's student at the same university and is currently working as a Data Analyst with expertise in ETL processes, dashboard designs, and database optimizations. His fields of interest are: data analysis, machine learning, deep learning, and LLM models.



## Detection of Fake News Using Deep Learning and Machine Learning

Gabriela CHIRIAC, Ada Maria CATINA

Bucharest University of Economic Studies

Faculty of Cybernetics, Statistics and Economic Informatics

Bucharest, Romania

[chiriacgabriela20@stud.ase.ro](mailto:chiriacgabriela20@stud.ase.ro); [catinaada20@stud.ase.ro](mailto:catinaada20@stud.ase.ro)

*Automatically identifying fake news is a complex challenge, involving detailed knowledge of how fake news is propagated and advanced data processing technologies. The use of Machine Learning and Deep Learning algorithms for misinformation detection involves a continuous learning process as manipulation methods are constantly evolving. Effective detection of fake news requires constant adaptation of algorithms to keep up with new disinformation methods. While these technologies offer promising solutions, the challenge is to calibrate them properly so that they work optimally in different contexts.*

*This paper explores automated methods for detecting fake news, analyzing the effectiveness of the various techniques and how they can be improved to face the challenges of data quality, domain variability and the continuous evolution of disinformation strategies.*

**Keywords:** *misinformation, Machine Learning, Deep Learning, fake news.*

### 1 Introduction

In the age of digitalization and social media platforms, the spread of fake news has become a real challenge. Today, online platforms provide us with access to information, but this freedom also comes with the risk of falling prey to cybercriminals who spread fake news.

Fake news is defined as text, images, videos or audio files intentionally created to misinform public opinion. Social media platforms enable misinformation to spread very quick, exposing thousands of users just in few minutes. Their purpose is to influence people's opinions about an individual or group of individuals, to create social unrest and to undermine trust in institutions.

Because of the quick spread and big volume of information, distinguishing between real and false news has become very difficult. Manually verifying information can be a slow and difficult process due to the large amount of data being produced every day.

Social media is a key way to consume news, but it can have both advantages and disadvantages. It is great for access, quick and cheap, as well, but it also makes the spread of fake news easier. In many cases, the quality of information can be unreliable, as some information are intentionally falsified. Detecting this type of deception has become a research topic, attracting increased interest in recent years [1]. According to one of the EU Agency for Cybersecurity reports [2], disinformation and misinformation are a critical challenge to global security. In this report we can find that the use of cloud computing technologies and AI algorithms contributes to the creation of misinformation, accentuating challenges related to the *security and integrity of the online environment*.

Another important aspect highlighted by the European data Protection Supervisor [3] is the use of complex applications called *bots*, which are organized into networks and used heavily to amplify disinformation content. Bot networks are organized by foreign actors attempting to



mask the true initiators, thus further complicating the detection process. However, recent research indicates that while automated bots contribute to the spread of fake news, the influence of human behavior, especially *word of mouth* marketing, has an even greater impact. As such, the issue of how to combat the spread of fake news can't come down to how to identify the people who produce and spread it. We must equip people to spot fake news and teach them critical thinking so they can keep the danger away. Thus, media education becomes the key factor in the fight against fake news, helping to build a society that is better prepared to deal with manipulation [3].

This phenomenon has serious consequences for society, and various advanced methods have been developed to combat it, applying machine learning and deep learning algorithms. The algorithms are trained to verify news content by exposing them to large volumes of data. This training process involves the algorithms learning to distinguish between real and fake news based on a set of properties extracted from the text, images, or metadata associated with the news. Since the spread of fake news is constantly changed, these methods need to continuous update to keep up with sophisticated tactics used by disinformation creators to mislead both users and detection systems. [3]

This paper aims to explore advanced *Machine Learning and Deep Learning* methods used in the automatic detection of fake news. This study will examine different techniques and algorithms that can be applied to classify informational content, aiming to evaluate their effectiveness in detecting fake news and combating them.

## 2. Literature Review

Numerous studies have explored solutions for identifying and classifying fake news from various data sources, using different

approaches to increase the accuracy of predictions. In this section, we will analyze significant contributions from the research literature that address various techniques and methods for detecting this problem.

In 2015, several researchers [4] adopted a text-based approach, dividing fake news into three typologies: serious inventions, large-scale jokes, and parodies. They explore the difficulties encountered in developing a false news detection system, highlighting the importance of taking into account the typology and context in which they are generated. Perez-Rosas and other researchers [5] have developed an automated algorithm that combines lexical, syntactic and semantic information to detect fake news, bringing more complexity and efficiency to the detection process. In a similar study [6] proposed a hybrid method for detecting fake news, which combines linguistic analysis with social networks evaluation techniques. This method has proven to be successful in identifying fake news, given that social network analysis can reveal the ways in which false information is propagated and amplified on online platforms.

In 2016, Dadgar and collaborators [7] applied feature-extracting techniques such as TF-IDF, and implemented machine learning algorithms such as SVM (support Vector machines) to classify fake news into various categories. Ruchansky, Seo, and Liu [8] proposed a hybrid algorithm named CSI, which integrates three essential components: capture, score, and integration to improve predictions about fake news. This approach underlines the importance of integrating information from multiple sources to build a more robust detection system.

Regarding social network-based techniques, Shu, Wang and Liu [9] studied another model that analyzes factors such as the position of news and user interactions on social networks. This approach is extremely relevant, given how much fake

news and manipulation rely on social media platforms. Similarly, a study by Jin, Cao, Zhang, and Luo [10] proposed an approach based on identifying conflicting points of view in social networks. By applying this method to real data sets, the authors demonstrated how differences of opinion can be indicators of the presence of misleading information.

Several researchers participating in a conference on natural language processing [11] used an attention-based long short-term memory (LSTM) network to detect fake news, demonstrating the use of advanced technologies, such as neural networks, to improve the detection process.

Recent studies, such as that of Janze Christian [12], have applied these techniques in the context of the 2016 US election to observe the impact of fake news on election campaigns. Butain and Golbeck [13] have created an automated system that detects fake news on Twitter, a platform where its dissemination is a major problem.

To reduce the impact of fake news, several researchers [14] have proposed a competitive model that analyzes the relationship between the original and updated false information, with the aim of minimizing its effect on the public.

On the other hand, Tschischek and collaborators [15] proposed a method based on the human signals, using Bayesian inference to improve detection accuracy, and Guacho, Abdali and Papalexakis [16] introduced a semi-supervised false news detection technique, which combines human signals with machine learning algorithms.

These innovative research and solutions highlight the complexity and diversity of approaches in the field of detecting fake news. While technologies have advanced in detection of fake news, the challenges remain due to the quick spread of fake news and to the development of new methods and tools used for disinformation.

### 3. Methodology

This paper is based on a comparative literature review methodology, aiming to highlight the most efficient methods for automatic detection of fake news using Machine Learning and Deep Learning.

The study was conducted by identifying and selecting relevant scientific papers, published in specialized journals and conference in the field of artificial intelligence and natural language processing. This selection criteria included topic relevance, types of algorithms used, datasets used and evaluation metrics reported. The analysis covered both classical models, such as Logic Regression, SVM, Random Forest, XGBoost, KNN, and advanced deep learning models, such as CNN, RNN, BERT. The models were compared according to the performance obtained in the analyzed papers.

### 4. Evaluation Metrics

Various metrics are used to evaluate the performance of algorithms, most of which are based on the confusion matrix. The confusion matrix is a table that shows and compares the actual values with the predicted outcomes of a classification model (Table 1).

**Table 1.** Confusion matrix

		<i>Current values</i>	
		1	0
<i>Predicted values</i>	1	True positive (TP)	False positive (FP)
	0	False negative (FN)	True negative (TN)

*Accuracy* measures the percentage of correctly classified instances out of the total number of instances analyzed. It is calculated as the ratio between the sum of True Positives and True Negatives and the

total number of instances. This indicator is particularly relevant in situations where the distribution of classes is balanced, when the number of examples in each category (e.g., fake news and real news) is approximately equal [17].

*Precision* indicates the proportion of instances correctly classified as positive, relative to the total number of instances predicted as positive. It is calculated as True Positives / (True Positives + False Positives). This indicator shows how accurate the positive predictions are and to what extent the model avoids misclassifying negative instances as positive [17].

*The Recall* (or True Positive Rate) measures the proportion of correctly identified positive instances in the total existing positive instances. It is calculated as True Positives / (True Positives + False Negatives). This indicator reflects the model's ability to detect all positive cases in the dataset, being especially useful when it is important to minimize omissions [17].

*The F1 Score* is the harmonic mean of precision and recall and gives a balance between precision and recall. This metric is particularly useful in unbalanced datasets, as it takes into account both the precision and the model's ability to detect all positive instances. In binary classification, True Positives (TP) is the number of events correctly classified as positive and False Positives (FP) is the number of events incorrectly classified as positive, True Negatives (TN) is the number of events correctly classified as negative and False Negatives (FN) is the number of events incorrectly classified as negative [17].

## 5. Machine Learning in detecting fake news

Machine learning is a branch of artificial intelligence which is concerned with the development of systems that can learn and adapt based on the data that they process. Algorithms are capable of generating

predictions, recognizing patterns in data, and making decisions based on the information extracted. They are classified into supervised, unsupervised, and semi-supervised machine learning algorithms, with the first two categories being used most frequently [18].

*Supervised learning algorithms* are trained on labeled datasets, where the results are known, in order to be able to correctly estimate future events. They learn to correctly associate input data with expected outcomes so that they can make accurate predictions based on new datasets.

**Logistic Regression** is a binary classification algorithm that estimates the probability that an instance belongs to one of two classes (in this case, true or false). This model is often used for simple classification tasks, it predicts whether an item is true or false based on features extracted from the text. Logistic Regression gives efficient results when there is a linear relationship between the input features and the output label. Although this model is not extremely complex, it is often efficient for data with simple relationships and has been used as a benchmark for comparisons with other more advanced algorithms. The mathematical functions of the hypothesis of logistic regression and cost to obtain an optimal probability are represented as follows:

$$h_{\theta}(X) = \frac{1}{1+e^{-(\beta_0+\beta_1 X)}} \quad [19]$$

$$\text{Cost}(h_{\theta}(x), y) = \begin{cases} \log(h_{\theta}(x)), & y = 1 \\ -\log(1 - h_{\theta}(x)), & y = 0 \end{cases}$$

Support Vector Machines (SVM) is a popular algorithm for classification problems, using a mathematical concept called *maximum edge*. It aims to find a hyperplane that separates the data in the two classes (true/false) by as large a margin as possible. SVM maximizes the

distance between the separating hyperplane and the nearest data points in each class (called support vectors). This large margin improves the generalization of the model. For cases where the data is not linearly separable, SVM uses a technique called kernel trick, which transforms the data into a higher dimensional space where linear separation becomes possible.

**Random Forests** are an ensemble of decision trees that help improve performance and reduce the risk of overtraining. Each tree in the forest is trained on a random subset of the dataset, obtained by the bootstrap method. At each node, a random subset of features is selected to determine the splitting criterion. The final prediction of the Random Forest model is made by majority vote (for classification) or by averaging the outputs (for regression) of all trees in the forest.

**K-Nearest Neighbors (KNN)** is an instance-based algorithm that classifies an article based on the nearest points in the dataset. KNN has been used in various studies to learn patterns of classification of fake news based on features extracted from the text. This is a simple and efficient algorithm, but it can be slower for large datasets because of the need to calculate distances between instances. Distances between two points can be calculated using the following formulas [19]:

$$\begin{aligned} \text{Euclidean distance} &= \sqrt{\sum_{i=1}^k (x_i - y_i)^2} \\ \text{Manhattan distance} &= \sum_{i=1}^k |x_i - y_i| \\ \text{Minkowski distance} &= \left( \sum_{i=1}^k |x_i - y_i|^q \right)^{1/q} \end{aligned}$$

**AdaBoost** (Adaptive Boosting) and **XGBoost** (Extreme Gradient Boosting) are two of the most popular boosting methods used to enhance the performance of

classification and regression models. Boosting is a technique that combines several weak models (usually simple decision trees) to build a strong and robust model. The basic idea is that the models are trained sequentially, with each new model focusing on correcting errors made by previous models.

AdaBoost increases the weight of instances that have been misclassified by previous models, making them more influential in the next iteration. XGBoost, on the other hand, applies a more advanced, gradient-based boosting technique that trains decision trees in a highly efficient and parallelized manner. In a study [19], XGBoost had a significant impact on improving overall accuracy and performance due to its ability to manipulate complex data and reduce prediction errors. XGBoost was one of the algorithms that contributed to the outstanding results for fake news classification.

**Decision Trees (CART)** is an intuitive machine learning method that builds a tree-like model, where each internal node represents a feature-based question, each branch corresponds to a possible answer, and the leaves indicate a classification or a numerical value (in the case of regression). In classification tasks, the tree splits the data based on features to create subsets that are as “pure” as possible. Splitting decisions are typically based on measures such as entropy or the Gini index, which assess the quality of the partitions.

In the research [19], decision trees (CART) have been used as part of an ensemble, playing a significant role in creating accurate classifications, especially for categorical or textual data. Although decision trees are easy to interpret, they can be less effective in handling data with multiple meanings or in detecting complex relationships.

A study conducted by two researchers from India [20] compared the performance of Logistic Regression and SVM in

detecting fake news distributed on social platforms. The results of applying the SVM and LR models are:

**Table 2.** Results from applying SVM and LR models on a fake news dataset [20]

Model	Accuracy	Precision
SVM	0.91	0.89
LR	0.95	0.93

Model	Recall	F1-Score
SVM	0.73	0.75
LR	0.79	0.83

On a dedicated dataset, LR achieved 95.12% accuracy and 93.62% precision, while SVM had 91.68% accuracy and 89.20% precision, showing a slight superiority of LR. The authors proposed a novel framework called Novel Fake News Detection (NFND) based on Logistic Regression and emphasized the need to extend the datasets and integrate advanced technique such as POS tagging, word2vec, and topic modeling to improve the performance.

In another study, some researchers [5], created two distinct datasets: one generated by crowdsourcing, covering six diverse domains (sports, business, entertainment, politics, technology and education) and another collected from the web, focusing on celebrity news. The true news articles originated from trusted sources such as ABCNews, CNN and The New York Times, while the fake articles were written by trained workers on Amazon Mechanical Turk to mimic journalistic style. For classification, a linear SVM was used with five-fold cross-validation, and the model performed exceptionally well using linguistic features such as readability, punctuation and the LIWC lexicon. The results showed high accuracy, sometimes even outperforming human annotators, although cross-domain generalization remained difficult, suggesting that structural and linguistic differences exist

between fake news across different content areas.

On the other hand, Conroy and collaborators [6] proposed a complementary approach, integrating SVM into a hybrid system that combines linguistic technique with social network analysis. In this context, SVM was applied to datasets containing both real and fake articles, including satirical and manipulated news, and demonstrated strong performance in identifying deceptive patterns. The authors concluded that this integration significantly improves fake news detection by leveraging both textual content and social context.

Both studies confirm the usefulness of SVM in automatic detection, as an efficient classification model based on linguistic features and as a part of complex hybrid system that combine different typed of data.

According to the study by Ahmad and others [19], the performance of several machine learning algorithms on four different datasets were evaluated. The obtained results were analyzed in order to compare the efficiency of each algorithm, depending on the characteristics of the datasets and their typology. The datasets used are open source and include both real and fake articles from various domains.

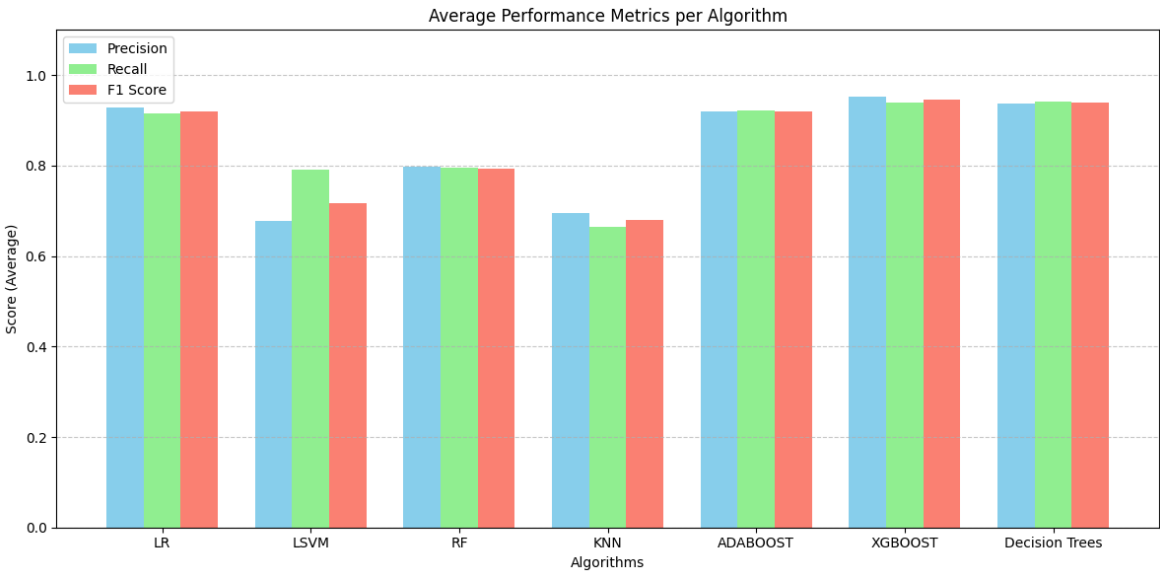
The first dataset, DS1 (ISOT Fake News Dataset), contains almost 45.000 articles, half are real, mainly from Reuters.com, and half are fake, sourced from disinformation websites, mostly political. The second dataset, DS2, available on Kaggle, includes more than 25.000 articles from various domains, split into training and test sets. The third dataset, DS3, also from Kaggle, has 3.300 articles from trusted sources such as CNN and The New York Times, as well as from untrusted sources focusing on sports, entertainment and politics. To allow for a more comprehensive evaluation, a combined dataset (DS4) was created, integrating all the articles from the previous sets.

Table 3 summarizes the average performance scores calculated across all four datasets to compare the overall effectiveness of the evaluated classification algorithms.

Figure 1 is a graphical representation of the average performance of the algorithms on all datasets [19].

**Table 3.** Average performance of learning algorithms [19]

Model	Average Precision	Average Recall	Average F1-Score
LR	0.93	0.92	0.92
LSVM	0.68	0.79	0.72
RF	0.80	0.80	0.79
KNN	0.70	0.67	0.68
AdaBoost	0.92	0.92	0.92
XGBoost	0.95	0.94	0.94
Decision Tree (CART)	0.94	0.94	0.94



**Fig. 1.** Average performance of learning algorithms [19]

XGBoost ranked as the best performing algorithm, achieving an average F1 score of 0.95. Its superior performance is due to its ensemble boosting method and advanced regularization techniques that minimize classification errors systematically. AdaBoost and Decision Trees also demonstrated strong results, with F1 scores of 0.92 and 0.94 respectively, benefiting from adaptive learning mechanisms and hierarchical decision structures [19]. Logistic Regression maintained a constant

performance (F1-score = 0.93), showing a particular efficiency on homogeneous datasets, while it showed a moderate decrease (to 0.87) on more heterogeneous data. This suggests its reliability for linearly separable problems, but also limitations in handling complex nonlinear relations. The analysis revealed suboptimal performance for linear SVM (F1 score = 0.73) and KNN (F1 score = 0.68). Linear SVM demonstrated a significant compromise between precision (0.68) and recall (0.79), while KNN performance was

affected by sensitivity to data noise and dimensionality [19].

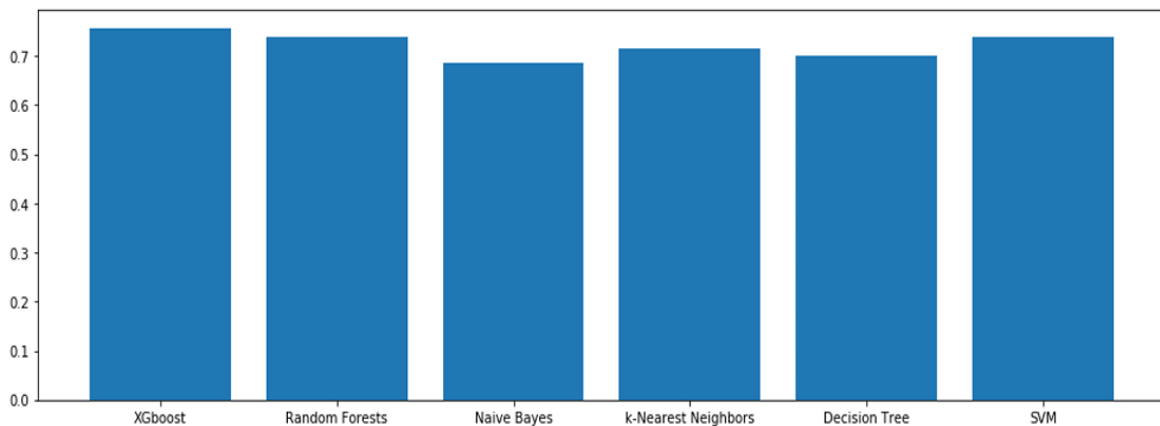
A relevant example of the application of machine learning algorithms is presented in the study [21], published in the IOP Conference Series: Materials Science and Engineering. The authors used the LIAR-PLUS Master dataset, which contains veracity-labeled claims from the politifact.com platform. The dataset was preprocessed using the NLTK and SAFAR v2 libraries, applying operations such as text cleaning, tokenization, POS tagging and linguistic features extraction (e.g. average word length and adjective frequency). The study compares the performance of several classification algorithms, including Naïve Bayes, Random Forest, XGBoost, Support Vector Machine (SVM), K-Nearest Neighbors

(KNN), Decision Tree and Linear Regression.

The experimental results showed that XGBoost algorithm achieved the best performance, with an accuracy of 75.30%, an F1-score of 77%, a precision of 76% and a recall of 0.75%. It was followed by SVM, which obtained an accuracy of 73.20% and Random Forest with an accuracy of 72.50%.

Simpler models, like Naïve Bayes and KNN, performed worse, with accuracies of approximately 65% and 62%, while Linear Regression gave modest results, confirming its limitations when addressing complex text classification tasks.

Figure 2 is a graphical representation of the average results of all the algorithms [21]:



**Fig.2.** Accuracy Results of all the Algorithms

In addition to individual model testing, the authors also explored hybrid models by combining multiple classifiers through majority voting strategies. The results indicated an improvement in performance: the hybrid model composed of XGBoost, SVM and Random Forest achieved an accuracy of 81.20%, demonstrating increased robustness and reduced classification errors.

The comparative analysis underscored the importance of selecting an appropriate

algorithm and complementary models can lead to superior performance in fake news detection tasks.

While supervised learning algorithms have shown promising results in the automatic detection of fake news, several studies have also highlighted their limitation, particularly regarding generalizability in novel contexts or across varied domains. For this reason, recent research has explored complementary methods that combine linguistic analysis and structural information.



*Algorithms with unsupervised learning* are trained on unlabeled datasets, where the correct results are not known. Their goal is to identify hidden patterns, structures or relationships in the data without knowing a previously provided output. These algorithms learn to organize and classify data on their own, discovering clusters or similar features, which can then be used to understand and interpret new data. This approach is useful when manual labeling is impossible. Although less precise than supervised methods, clustering provides valuable information in the absence of labels.

A key factor in one of the success factors behind machine learning algorithms is *parameter tuning*. This means that you will have to tune important parameters like the number of trees in a random forest, the level of regularization in a logistic regression or the number of neighbors in the KNN. The optimal choice of these parameters allows algorithms to become much more efficient in identifying false news, especially when using complex and large datasets.

Feature selection techniques, such as dimensionality reduction (e.g. PCA), can help improve results by eliminating irrelevant variables and preventing over-training. Appropriate adjustments of these parameters and features can lead to better performance, increasing the effectiveness of systems designed to detect fake news.

A relevant example of the application of unsupervised learning in fake news detection is presented in the study by Yang and colleagues [22], where they develop an original method based on probabilistic modelling. This research tackles the challenge of identifying fake news without using manually labeled datasets, which differentiates it from most existing approaches. The proposed model, called **UFD** (Unsupervised Fake News Detection), is based on the idea that users' interactions on social media, such as likes, retweets or comments on posts, may reflect

their perceptions of a news story's veracity. They consider both news veracity and user credibility as latent variables and incorporate them into a generative model built on Bayesian networks. To validate the method's performance, they used two real datasets: LIAR and BuzzFeed News.

The UFD model demonstrated notably strong results. Compared to other unsupervised methods (Majority Voting, TruthFinder, LTM, CRH), the proposed model performed better.

For **Majority Vote** each news item gathers the opinions of verified users, and the most common version is considered true.

**TruthFinder** tries to find out which information is true, even if user opinions contradict each other. It analyzes the conflicts between opinions and calculates which information is most likely to be correct, even without knowing in advance what is true or false.

**LTM** is a model that recognizes that users can be wrong. It tries to discover the truth even if some people give wrong information, but it needs a simple structure in which each source says something about particular news item.

**CRH** assesses how correct each user is in general. If a user has given correct information many times, their opinion will count more in the final decision. This calculates how credible everyone is and determines which information is true.

On the LIAR set, UFD achieved an accuracy of 75.90%, while the other methods ranged between 58% and 64%. Additionally, the model attained an F1-score of 74.1%, exceeding the performance of the majority voting method of over 20 percentage points.

An important advantage of this approach is its dual capability: it not only estimates the veracity of news content but also assesses the credibility of users. This dual output enhances the model's reliability and its

capacity to extract valuable insights from unstructured data.

Machine Learning has shown considerable potential in detecting fake news, with choosing the right algorithms being the most important step in obtaining accurate results. Algorithms such as XGBoost, AdaBoost and logistic regression have proven effective in correctly classifying articles, while unsupervised learning can uncover hidden patterns in the data, bringing additional value to the detection process. Combining multiple algorithms can also help create a more robust model.

By integrating multiple algorithms and evaluating them across diverse datasets, it becomes possible to identify the most suitable models for the automatic detection, depending on the type of data, the complexity of the relationships between variables and the computational resources available. The reviewed studies confirm that ensemble models, such as XGBoost and Random Forest, offer an excellent balance between accuracy and robustness, like Logistic Regression, and can also yield competitive results in more constrained scenarios, with the added benefit of interpretability. Moreover, the adoption of hybrid models that combine linguistic features, contextual signals and meta-information extracted from social media represents a promising direction for enhancing the reliability of fake news detection systems.

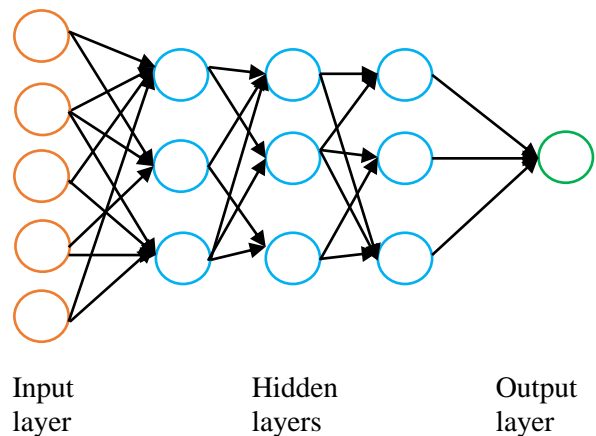
However, the ultimate success of a system is highly dependent on continuous optimization of algorithms and hyperparameter tuning, which can make the difference between an accurate model and a less performing one. Future research should incorporate these aspects to increase the accuracy and adaptability of counter-disinformation systems.

## 6. Deep Learning in fake news detection

Another branch of Artificial Intelligence, related to Machine Learning, is *Deep*

*Learning*, abbreviated DL. Deep Learning consists of learning methods that allow computers to learn independently, without the intervention of a human factor to define rules or knowledge. These models are structured as an artificial neural network because of their architecture which is made of interconnected nodes across multiple layers, similar to the neurons of a biological brain. The difference between DL and ML is that the latter involves the use of neural networks that have an input layer of neurons, an output layer and sometimes 1-2 hidden layers, whereas Deep Learning uses *deeper* neural networks, as the name suggests, with multiple hidden layers.

An example of a deep neural network can be seen in the figure below:



**Fig. 3.** Example neural network with multiple hidden layers

Natural Language Processing, or **NLP**, is another branch of AI, like ML and DL. Using NLP, computers can understand human language [23]. This technology is the basis of intelligent assistants. With NLP, these assistants can understand and reproduce the human language. NLP provides a number of data *pre-processing techniques* that are essential for ML and DL. These methods necessary especially for the analysis of a group of sentences, such as a news story. An NLP method is

**tokenization** in which a text is split into a sequence of tokens, consisting of words or parts of words. Another approach is the *Bag-of-Words* model, in which a data set is viewed as a multitude of words put together, ignoring their order. This calculates the frequency of occurrence of words in the text. One use of Bag-of-Words is searching for certain information on the Internet or other search engines. Bag-of-Words technique along with **TF-IDF** are vectorization methods often used in ML and DL models for fake news detection.

Another vector-based NLP technique is the **word embeddings** technique, which assigns different real number vectors to words, with the resulting representations being distributed, with a reduced dimension. This method is considered a more improved version of Bag-of-Words, the vector space being continuous and multidimensional. Models often used in fake news detection, pre-trained with word embeddings, are *GloVe* and *Word2Vec*. These have the ability to train very large datasets. *Word2Vec* represents performance in detecting syntactic relations between words, while *GloVe* does better with global semantic relations.

A first and common DL model is **CNN** (*Convolutional Neural Network*), first introduced by Kunihiro Fukushima [24]. Convolutional Neural Network implies the use of convolutional layers within the model. These layers filter the input data so that only relevant and useful information is extracted. CNN is basically used for image classification and recognition as input data. Convolutional layers may have other pooling layers attached to them. These layers reduce the size of images to decrease the number of parameters in a network. With other words, a pooling layer is a small portion of the input image and the convolutional operation helps extract its most important features.

A study by M. F. Mridha [25] highlights that *max pooling and average pooling* are

the most commonly used functions in CNNs. Their research identified best parameters values for a CNN model in order to perform in detection of the fake news: the dense layer has 100 units, there are 100 filters, and the filter size is 5. *GlobalMaxPooling1D*, which performs global pooling, has the highest score among the CNN approaches, making it the most effective solution for detecting fake news.

As with any other learning model, certain problems arise, such as error reduction or *overfitting*, which occurs when model performance is poor on new data. One of the solutions to the overfitting problem is *regularization*, which is most common in detecting false news, according to Mridha's research.

Another often used model of Deep Learning is **RNN** (*Recurrent Neural Networks*), used for processing sequential data, such as speech and language. The structure of these neural networks includes recurrent links that *memorize* the information from previous execution of the same computation process. In this way, the final result depends on the results obtained during the whole process [24]. A problem encountered in the RNN application is the *disappearance of the gradient in time*.

A variation of the RNN model is RNN with Long Short Term Memory (**LSTM-RNN**) [17]. The added memory (LSTM) is intended to retain information from previous computations over a longer period. In detecting fake news it is important to understand the context, to look at the information as a whole for a correct classification, that is why the LSTM-RNN model is often used in this case. Nowadays, most of the fake news are appearing on social networks. In their study, Sahoo and Gupta [25] tried to detect such news items on the Facebook platform. They considered both the data of the posts and the account information of the people who published them, applying the LSTM-

RNN model. However, due to too much data, the runtime was too long.

Another variation of the RNN model is **GRU** (*Gated Recurrent Unit*), similar to LSTM, but with a simpler architecture. GRU needs fewer parameters, making the train process more efficient. The main difference between GRU and LSTM model is how they manage the memory. LSTM has a separated memory part, which is updated through three gates (input, output and forget). On the other hand, GRU combines the memory and the hidden parts into one, using only two gates (reset and update). Through these gates, GRU model chooses what information is modified. At a conference from 2024, Elfaik and Nfaoui [26] did a comparative study between GRU and other Deep Learning methods, LSTM and RNN, used to detect fake news. They used ISOT Fake News Dataset and after cleaning the data and applying selected methods to identify the fake news, they observed that GRU model had a higher accuracy compared to the others, as we can observe in Table 4 bellow. This means that GRU model is highly effective for detecting fake news with minimal manual features extraction.

**Table 4.** GRU vs. LSTM and RNN – accuracy values [26]

Model	Accuracy
LSTM	0.9969
RNN	0.7448
GRU	0.9983

The LSTM of a recurrent neural network can also be bidirectional, related to the **BiLSTM-RNN** model. This is a variant of the traditional RNN model, but which provides a memory that allows data to be processed from the beginning to the end and from the end to the beginning. BiLSTM-RNN is used when both the past and the future need to be analyzed, it is

more powerful than the classical LSTM, but also more expensive.

One pre-trained DL model is the **BERT** model, short for Bidirectional Encoder Representation from Transformers. This model is based on transformers, as the name suggests, and was introduced by Google in 2018 [17]. Transformers have the ability to process a word by taking into account its relationships with other words in the text, allowing the model to understand its context. Thus, BERT is used for natural language processing (**NLP**) in many situations, like translating languages.

As with RNN, BERT has several variations, one of which is **ALBERT**, which is derived from the *A Lite BERT model*, considered to be efficient to use for false news detection. In contrast to BERT, ALBERT uses a smaller projection layer and applies the *Weight Sharing principle*, which divides the weights among all layers, thus reducing the number of parameters.

**RoBERTa** (Robustly Optimized BERT Approach) is another improved variant of the BERT model, introduced in 2019 by Facebook AI researchers. The difference between the two models is that RoBERTa can process a much larger volume of data and uses improved training procedures. This makes RoBERTa a more powerful tool to use, especially for fake news detection. However, in 2020, following a study, the **FakeBERT** model specifically designed for fake news detection was introduced. FakeBERT is based on the architecture of the BERT model, and the difference costs in the training set containing fake and real news [27]. Another specific feature of FakeBERT is the use of the back-translation technique. This technique involves translating a real news text into a language, and then translating it back into the original language, generating synthetic data that is added to the training set.

## 7. Deep Learning vs. Machine Learning in fake news detection

Machine Learning and Deep Learning methods show performance in identifying fake news, as we can observe in this paper. ML utilizes classical algorithms such as SVM or Random Forest, while DL uses the principle of deep neural networks such as CNN and RNN. In contrast to ML, DL models have the ability to process large datasets containing raw texts, without human intervention, learning automatically. However, due to its higher performance, Deep Learning involves higher costs in terms of resources and processing time. In contrast, traditional models can be trained and deployed much faster on modest hardware, making them a more practical solution in low-resource environments.

Both branches of Artificial Intelligence have performed well in fake news detection, but their performance differs depending on the details of the problem. For example, if one wants to classify a text as fake or not, then an ML model can be used, but if one wants to analyze a large dataset, such as all the news posted on a web page, then an DL model is better to use, due to its improved ability to work on large datasets.

Several studies have been conducted over time comparing different ML and DL models applied on the same datasets. The Deep Learning methods presented earlier in this paper are the most common methods used in different studies on fake news detection. An example of such work is the comparative study of Deep Learning and Machine Learning methods applied to identify fake news, published in the *Carol National Defense University Bulletin* [17]. The datasets underlying the research include fake news from ISOT, BuzzFeed and PolitiFact websites. DL models, also described in this paper, were applied and results obtained as follows:

**Table 5.** Results from applying DL models on a fake news dataset [17]

Model	<i>Accuracy</i>	<i>F1 Score</i>
RoBERTa	0.99	0.99
LSTM-RNN	0.96	0.97
BiLSTM-RNN	0.98	0.97
ALBERT	0.97	0.97
FakeBERT	0.98	0.99
BERT	0.98	0.98
CNN	0.96	0.96

In the table above, we can observe that after applying the models, the best performing model in identifying fake news is RoBERTa, with accuracy and F1 score values very close to 1 and 0.99, respectively. The difference between RoBERTa and the other models is small, which shows that all models were able to correctly identify almost all fake news, the worst performing model being CNN, with an accuracy and F1 score of 0.96.

A case study is the work of S. Repede [17], mentioned earlier in the paper. He applied on 4 datasets several machine learning models. Analyzing the evaluation metrics, it was found that the most efficient model, with the highest performance, is the *RoBERTa* model, with values above 0.99, very close to the maximum value 1.

Another research conducted by Arshad Ali and Maryam Gulzar [28] was based on the detection of fake news on social networks related to the COVID-19 pandemic. In this research, the authors tried to combine two machine learning models, **BERT** (DL) and **SVM** (ML), along with an evolutionary algorithm, **NGSA-II** (Non-dominated Sorting Genetic Algorithm II), to obtain better results, thus creating a new hybrid model. After preprocessing the COVID-19 news dataset, different traditional ML and DL models were applied, including:

**Table 6.** Results applying classical ML and DL models on the COVID-19 dataset [28]

Model	Accuracy	Precision
SVM	0.60	0.65
Random Forest	0.68	0.63
K-Neighbour	0.64	0.69
BERT	0.63	0.72
CNN	0.72	0.73
Model	Recall	F1 Score
SVM	0.71	0.55
Random Forest	0.7	0.66
K-Neighbour	0.65	0.7
BERT	0.63	0.67
CNN	0.82	0.77

We can observe that the applied models have an average performance, the best being CNN, with the highest evaluation metrics values. After the application of the proposed hybrid model, BERT+NSGA-II+SVM, it was found that it performs much better than the traditional models applied individually. The F1 score (0.83) and accuracy (0.8) have values above 0.8, recall has a value of 0.9, and precision is 0.76. All metrics have higher values than those of the CNN model, the best performing of the traditional models applied.

At a conference, Chang [29] conducted a study on different Machine Learning and Deep Learning algorithms used to combat fake news. He used the ISOT dataset containing fake and real news. After pre-processing it, the author compared 15 algorithms, and the results showed that Deep Learning models performed the best, especially the BERT model with 99.95% accuracy. In second and third place were the BiLSTM and LSTM models, showing the superiority of Deep Learning on this dataset. Regarding Machine Learning, the best result belongs to the SVM model, with an accuracy of 98.65%.

A comparative study was carried out by Kishwar and Zafar [30] based on a dataset of news about Pakistan. The two researchers used Google Fact Checker along with a series of words to extract relevant news. In addition to Google Fact Checker, data was also extracted from other sources such as Kaggle. After applying different Machine Learning models and technologies, Deep Learning was found to perform better in identifying fake news. The CNN model in conjunction with GloVe achieved an F1 score of 0.93, while LSTM scored 0.94. An interesting thing introduced in the study is a questionnaire completed by 57 people who had to categorize 10 news stories as fake or real. The results of the questionnaire showed that most of the fake news stories correctly identified by those individuals were also classified as fake by the models applied. However, there were many misclassifications of real news stories as fake because they had a similar writing style to fake news, with information that appeared to be false. Thus, it was shown that ML models can overcome human judgment.

Similar to the previously mentioned studies, Alghamdi, Lin and Luo [31] did a study in which, on several fake news

datasets, LIAR, PolitiFact, GossipCop and COVID-19, they applied different models of both ML and DL. LIAR is a rather voluminous dataset used for identifying fake news, composed of US political statements. Similarly, PolitiFact is a dataset of political statements available on the website of with the same name. On the other hand, GossipCop contains data on celebrity news stories, and COVID-19, as the name suggests, contains news about the COVID-19 pandemic, both sets categorized as real or fake. Their study becomes even more interesting because they applied the same model several times, each time using different pre-processing techniques. Their results showed that the classical ML models, applied together with TF-IDF, perform best on the LIAR dataset, compared to more advanced models such as DL or hybrid models. In contrast, on the PolitiFact dataset, RoBERTabase performs best with a high F1 score of 0.93. Regarding the GossipCop dataset, classical ML models performed better, while on the COVID-19 dataset the BERTbase model, belonging to Deep Learning, performed best. This study shows that there is no universal ideal method that gives the same performance on all datasets, as context as well as other factors (resources, data volume and so on) matter a lot.

## 8. Conclusion

The problem of the spread of fake news today is growing. People spend much of their time on social media platforms like Facebook, Instagram and TikTok. For various reasons, certain individuals or *bots* post various news, ads or videos containing false information with the aim of mass manipulation or monetization. Therefore, the use of a mechanism to identify false information is essential. Both Machine Learning models, such as Random Forest and SVM, and Deep Learning models, such as RNN and FakeBERT, can be used for this purpose, as we have seen in this paper.

Machine Learning and Deep Learning perform well on this false news topic. However, depending on the size of the dataset, the complexity of the situation and the available resources, an ML model may be more efficient than a DL model or vice versa. It has been shown that for large datasets, in particular fake news datasets, Deep Learning models perform better due to their LSTM, BiLSTM and BERT methods. However, technologies are evolving, and hybrid solutions that have appeared in recent years, such as the BERT+NSGA-II+SVM model proposed and analyzed by Ali and Gulzar [28], which combines ML models with DL models, NLP techniques and other technologies, perform much better and represent the future of combating online misinformation.

## References

- [1] K. S. A. W. S. T. J. & L. H. Shu, „Fake News Detection on Social Media: A Data Mining Perspective,” *ACM SIGKDD Explorations Newsletter*, vol. 19, nr. 1, pp. 22-36, 2017.
- [2] U. A. f. C. ENSIA, "ENISA THREAT LANDSCAPE 2022," *European Union Agency for Cybersecurity*, 2022.
- [3] D. P. Supervisor, „Fake News Detection,” 2023. [Interactiv]. Available: [https://www.edps.europa.eu/press-publications/publications/techsonar/fake-news-detection\\_en](https://www.edps.europa.eu/press-publications/publications/techsonar/fake-news-detection_en). [Accessed 27 3 2025].
- [4] Y. C. N. J. C. V. L. Rubin, "Deception detection for news: three types of fakes," *Proceedings of the Association for Information Science and Technology*, vol. 52, no. 1, pp. 1-4, 2015.
- [5] K. A. L. R. M. V. Perez-Rosas, "Automatic Detection of Fake News,"



- arXiv preprint arXiv:1708.07104, 2017.
- [6] V. L. R. Y. C. N. J. Conroy, "Automatic deception detection: Methods for finding fake news," *Proceedings of the Association for Information Science and Technology*, vol. 52, no. 1, pp. 1-4, 2015.
- [7] M. S. A. M. F. SM H. Dadgar, "A novel text mining approach based on TF-IDF and Support Vector Machine for news classification," *Proceedings of IEEE International Conference on Engineering and Technology*, pp. 112-116, 2016.
- [8] S. S. Y. L. N. Ruchansky, "CSI: A hybrid deep model for fake news detection," *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pp. 797-806, 2017.
- [9] S. W. H. L. K. Shu, "Understanding User Profiles on Social Media for Fake News Detection," *IEEE Conference on Multimedia Information Processing and Retrieval*, 2018.
- [10] J. C. Y. Z. J. L. Z. Jin, "News Verification by Exploiting Conflicting Social Viewpoints in Microblogs," *Thirtieth AAAI Conference on Artificial Intelligence*, vol. 30, no. 1, pp. 2972-2978, 2016.
- [11] Q. L. R. X. M. L. C. R. H. Y. Long, "Fake news detection through multi-perspective speaker profiles," *Proceedings of the Eighth International Joint Conference on Natural Language Processing*, pp. 252-256, 2017.
- [12] M. R. C. Janze, "Automatic detection of fake news on social media platforms," *Proceedings of 21st Pacific-Asia Conference on Information Systems*, 2017.
- [13] J. G. C. Buntain, "Automatically identifying fake news in popular twitter threads," *Proceedings of the IEEE International Conference on Smart Cloud, New York*, pp. 208-215, 2017.
- [14] W. H. C. J. F. G. L. H. Zhu H., "Information dissemination model for social media with constant updates," *Physica A: Statistical Mechanics and its Applications*, vol. 502, pp. 469-482, 2018.
- [15] S. M. G. R. A. M. A. K. S. Tschatschek, "Fake news detection in social networks via crowd signals," *Proceedings of the World Wide Web Conferences, France*, pp. 517-524, 2018.
- [16] S. A. E. E. P. G. B. Guacho, "Semi-supervised content-based fake news detection using tensor embeddings and label propagation," *Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp. 322-325, 2018.
- [17] R. B. Ş. E. Repede, "O comparație a modelelor de inteligență artificială folosite pentru detectarea știrilor false," *Buletinul Universității Naționale de Apărare »Carol I«*, no. 1, pp. 81-99, 2023.
- [18] Oracle, „Ce este machine learning?,” 2025. [Interactiv]. Available: <https://www.oracle.com/ro/artificial-intelligence/machine-learning/what-is-machine-learning/>. [Accesat 27 Martie 2025].
- [19] M. Y. S. Y. M. O. A. I. Ahmad, "Fake News Detection Using Machine Learning Ensemble Methods," *Complexity*, vol. 2020, no. 1, 2020.
- [20] M. A. N. Leela Siva Rama Krishna, "Fake News Detection System using Logistic Regression and Compare Textual Property with Support Vector Machine Algorithm," *Proceedings of the International Conference on Sustainable Computing and Data Communication Systems*, 2022.

- [21] N. A. H. S. M. R. Z Khanam, "Fake News Detection Using Machine Learning Approaches," *IOP Conference Series: Materials Science and Engineering*, 2021.
- [22] S. S. K. W. S. G. R. W. F. & L. H. Yang, "Unsupervised Fake News Detection on Social Media: A Generative Approach," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 1, pp. 5644-5651, 2019.
- [23] Eppright, „Ce este NLP (procesarea limbajului natural)?,” 2021. [Interactiv]. Available: <https://www.oracle.com/ro/artificial-intelligence/what-is-natural-language-processing/>.
- [24] O. Y. A. U. a. Y. D. M. Coskun, "An Overview of Popular Deep Learning Methods," *European Journal of Technic*, Vol. 7, Number 2, pp. 165-176, 2017.
- [25] J. K. M. A. H. M. M. M. a. M. S. R. M. F. Mridha, "A Comprehensive Review on Fake News," *IEEE Access*, pp. 156151-156170, 2021.
- [26] H. N. H. Elfaik, "Automatic Detection of Fake News Using Gated Recurrent Unit Deep Model," in *5th International Conference on Innovative Data Communication Technologies and Application*, 2024.
- [27] G. a. P. N. R. K. Kaliyar, „FakeBERT: Fake news detection in social media with a BERT-based deep learning approach,” *Multimedia Tools and Applications*, pp. 11766-11788, 2021.
- [28] M. G. A. Ali, "An Improved FakeBERT for Fake News Detection," *Applied Computer Systems*, Vol. 28, No. 2, pp. 180-188, 2023.
- [29] L. Chang, "Comparison of Machine Learning and Deep Learning Algorithms in Detecting Fake News," in *Proceedings of the 28th World Multi-Conference on Systemics, Cybernetics and Informatics*, 2024.
- [30] Z. A. Kishwar, "FakenewsdetectiononPakistaninewsusingmachinelearninganddeep learning," *Expert Systems With Applications*, vol. 211, pp. 1-10, 2023.
- [31] Y. L. S. L. J. Alghamdi, "A Comparative Study of Machine Learning and Deep Learning Techniques for Fake News Detection," *Information*, pp. 1-28, 2022.



**Gabriela Chiriac** graduated in 2020 from the Faculty of Accounting and Management Information Systems and she is currently a final year student in the master's program Databases – Business Support at the Faculty of Cybernetics, Statistics and Economic Studies, Bucharest University of Economic Studies. Professionally, she works as a Data Engineer, focusing on developing and maintaining ETL flows, managing databases and providing technical support for internal stakeholders.



**Ada Maria Catina** graduated in 2020 from the Faculty of Cybernetic, Statistics and Economic Studies, specializing in Economic Informatics. She is currently in her final year of the master's program Databases – Business Support at the Faculty of Cybernetics, Statistics and Economic Studies, Bucharest University of Economic Studies. Her academic interests focus on databases, data science and the application of machine learning techniques.

## Big Data and Business Intelligence: Synergies for Leveraging Competitive Advantage

Alexandru-Dan TURIAC, Mihai-Paul UNGUREANU, Mihnea-Cristian TACHE  
Bucharest University of Economic Studies  
Faculty of Cybernetics, Statistics and Economic Informatics  
Bucharest, Romania  
[turiacalexandru20@stud.ase.ro](mailto:turiacalexandru20@stud.ase.ro), [ungureanumihai20@stud.ase.ro](mailto:ungureanumihai20@stud.ase.ro),  
[tachemihnea20@stud.ase.ro](mailto:tachemihnea20@stud.ase.ro)

*In an increasingly competitive economic environment, the integration of Big Data technologies with Business Intelligence (BI) is a key factor in optimizing decision-making processes and enhancing organizational performance. Big Data enables the collection and processing of massive volumes of information from diverse sources, while BI facilitates their transformation into strategic insights through advanced analytical methods. This paper analyzes the technologies used in managing big data, the analytical and predictive models applied to extract value from data, as well as the benefits and challenges of implementing these solutions within organizations. Key aspects are highlighted, such as improved decision-making, operational optimization, and increased competitiveness, alongside difficulties such as high data volume, infrastructure complexity, and security and compliance requirements. The study emphasizes the importance of an integrated approach based on effective methodologies for fully leveraging the potential of Big Data and BI in the business environment.*

**Keywords:** *Big Data, Business Intelligence, analytical models, prediction, competitive advantage, decision optimization, digital transformation.*

### 1 Introduction

In a business environment marked by rapid digital transformation, an organization's ability to manage and capitalize on data is a crucial determinant of competitive success. The exponential expansion of data volumes—originating from sources such as commercial transactions, social networks, IoT sensors, and online interactions—necessitates the adoption of advanced technological solutions. In this context, the integration of Big Data systems with Business Intelligence (BI) provides new perspectives on decision-making processes, enabling the identification of trends, operational optimization, and increased organizational efficiency.

Big Data refers to the set of technologies and methods used to collect, store, and process massive volumes of structured and unstructured information, while BI focuses on transforming raw data into strategic

knowledge through complex analyses and visualization tools. The synergy between these two fields enables not only real-time access to relevant information but also the implementation of predictive models that enhance decision-making processes and contribute to the sustainable development of organizations.

This paper analyzes the technologies used in managing big data and integrating them with BI platforms, emphasizing processing methodologies and specific tools. It further examines the analytical and predictive models applied in various industries, along with the benefits and challenges related to adopting such solutions. Through this approach, the study aims to highlight the impact of digital transformation on organizations and to provide a perspective on how Big Data and BI can be strategically utilized to achieve long-term competitive advantage.

## 2 Big Data

In the digital era, big data has become a fundamental element in decision-making processes and organizational development strategies. The exponential increase in the volume of available data is fueled by various sources, such as social networks, IoT devices, financial transactions, sensor data, and online user interactions. This continuous expansion poses significant challenges related to storage, processing, and analysis, but also provides valuable opportunities for companies that manage to leverage it effectively.

Big Data is characterized by several key dimensions, initially known as the "3Vs"—volume, variety, and velocity—expanded with four additional dimensions: veracity, value, valence, and variability. **Volume** refers to the vast amount of data generated daily, exceeding the processing capacity of traditional database systems. As data volume increases, new challenges arise regarding storage methods and the infrastructures required to manage it. **Variety** refers to the diversity of data types collected, including structured data (relational databases), semi-structured data (XML, JSON files), and unstructured data (images, videos, social media text). This diversity requires advanced processing and integration methods to turn raw data into usable information. **Velocity** indicates the speed at which data is generated and processed, with some applications requiring real-time analysis. For instance, in industries such as finance or e-commerce, rapid data analysis can influence strategic decisions and optimize operational processes.

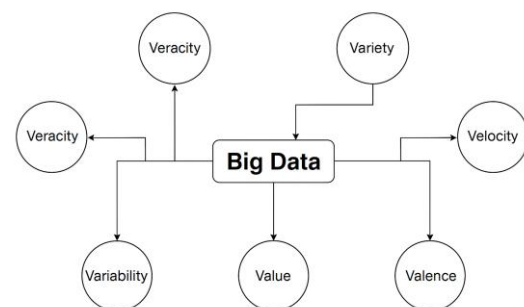
Beyond these fundamental dimensions, four more have been introduced for a more comprehensive understanding of Big Data. **Veracity** refers to the quality and accuracy of data, essential for making trustworthy decisions. **Value** reflects the benefits organizations can gain from data analysis, such as cost optimization, enhanced customer experience, or the development of new business models. **Valence**

describes the degree of connectivity between data sets. **Variability** refers to the dynamic and constantly changing nature of data, which is critical for correctly interpreting information.

Managing big data requires the use of advanced technologies such as distributed storage systems (Hadoop, Apache Spark), NoSQL databases (MongoDB, Cassandra), and parallel processing solutions. These technologies enable fast and efficient data analysis, providing companies with a competitive edge by enabling informed, data-driven decisions. [1]

### 2.1 ETL Processes (Extract, Transform, Load)

The digital transformation of industries has generated massive volumes of data from diverse sources such as social networks, IoT devices, transactional systems, and machine-generated logs. This phenomenon, known as Big Data, is characterized by a high volume of data, fast processing speed, a wide variety of data types, veracity, and value. Big Data holds tremendous potential to provide valuable insights and drive innovation, but its management and utilization come with significant technical and organizational challenges. At the core of addressing these challenges lies the ETL (Extract, Transform, Load) process, which involves extracting data from various sources, transforming it into a structured format suitable for analysis, and loading it into target systems such as data warehouses or data lakes.



**Fig 1.** The dimensions of Big Data

### **3 Traditional ETL Process Challenges**

Traditional ETL systems were designed in an era when data was relatively small, structured, and predictable. These systems relied on centralized architectures, batch processing, and predefined schemas to handle data integration tasks. In this context, the main challenges included optimizing query performance, managing changes in data schemas, and ensuring consistency between disparate data sources. However, the rise of Big Data has rendered these approaches insufficient, as traditional ETL processes can no longer keep up with the scale and speed required for handling massive and rapidly changing data.

#### **3.1 Challenges Introduced by Big Data**

With the emergence of Big Data, researchers and practitioners began exploring new approaches to ETL processes to address its unique characteristics. One significant research area has been scalability, particularly in the context of distributed processing frameworks such as Hadoop and Spark. These platforms allow for the parallel processing of large datasets, significantly reducing the time required for ETL tasks. However, using such technologies often requires substantial investment in infrastructure and expertise, which can be costly for smaller organizations.

Another key area of development has been real-time data processing. Traditional ETL systems, designed for batch processing, are not well-suited to the high-speed data generated by sources like financial transactions or social media. As a result, stream processing platforms such as Apache Kafka and Apache Flink have emerged as popular tools for managing real-time data, enabling organizations to process and analyze information as it is generated. Additionally, hybrid ETL systems that combine batch and stream processing have been developed to balance latency and throughput.

### **4 Data Variety and Associated Challenges**

Big Data frequently includes semi-structured and unstructured data such as JSON files, XML documents, images, videos, and text. Transforming and integrating these types of data is a major challenge, as most traditional ETL tools are optimized for structured data. Recently, emerging technologies in natural language processing (NLP), image analysis, and graph analysis have provided new methods for extracting and transforming unstructured data. Furthermore, schema-on-read approaches in data lakes have enabled more flexible data management, though this flexibility introduces greater complexity in terms of data governance.

#### **4.1 Data Quality and Governance in Big Data**

Data quality and governance remain persistent challenges in Big Data environments, as the volume and complexity of data make it difficult to ensure accuracy, completeness, and consistency. Various studies propose automated techniques for data quality assessment and anomaly detection, using statistical and machine learning methods. However, implementing these techniques at scale remains a significant hurdle. Therefore, governance frameworks are needed to balance flexibility and control, ensuring compliance with legal requirements and maintaining data trustworthiness.

### **5 The Role of Cloud Computing in Modernizing ETL Processes**

Cloud computing has played a transformative role in modern ETL practices, offering scalable and cost-effective solutions for processing Big Data. Cloud-based ETL tools provide elastic resources that can adapt to various workloads, reducing the need for upfront infrastructure investments. Integration with native cloud services, such as serverless

computing and managed databases, has further simplified the ETL process, allowing organizations to focus more on data analysis and less on infrastructure management. [2]

### 5.1 Storage (Data Lakes, Data Warehouses)

In the Big Data era, the immense volumes of data from sources such as social networks, IoT devices, and transactional systems require innovative storage and processing solutions. In this context, data lakes have emerged as essential systems for managing vast and complex datasets. A data lake is a centralized repository designed to store large amounts of raw data—structured, semi-structured, and unstructured. Unlike traditional databases, which require predefined schemas, data lakes allow data to be stored in its original format, providing maximum flexibility for analysis. Key features that define data lakes include their **remarkable scalability**, which allows them to efficiently manage vast volumes of data—often measured in petabytes—while remaining cost-effective, a fundamental requirement in Big Data contexts. Another essential characteristic is the use of a **schema-on-read** approach, meaning the data's structure is defined only when it is accessed, rather than at the point of ingestion. This provides a high degree of adaptability when working with heterogeneous datasets. Furthermore, data lakes are highly **flexible**, supporting a wide array of data formats such as JSON, XML, CSV, and various binary file types, making them well-suited for handling the diverse data sources typical in modern analytics. Moreover, they offer **low-cost storage** by leveraging distributed systems or cloud infrastructure, making them an economically viable solution for organizations managing large-scale data repositories.

### 5.2 Storage and Processing of Raw, Unstructured, and Semi-Structured Data

One of the main advantages of data lakes is their ability to store and process raw, semi-structured, and unstructured data originating from diverse sources. In the context of Big Data, such data is often generated by IoT devices, social networks, activity logs, or multimedia files. While traditional solutions like data warehouses are optimized for structured data, data lakes are specifically designed to handle large and varied datasets that cannot be efficiently managed by traditional systems. The data is commonly stored in distributed file systems (e.g., Hadoop Distributed File System - HDFS) or cloud-based storage solutions (e.g., AWS S3 or Azure Blob Storage). Processing is typically performed using parallel computing frameworks such as Apache Spark or Hadoop MapReduce, which allow for efficient analysis of massive datasets.

### 5.3 Common Use Cases

Data lakes are widely employed in scenarios that involve the analysis and processing of large and complex data volumes, which are typical of Big Data environments. One major application is in **Big Data analytics**, where organizations examine extensive datasets to uncover patterns, understand customer behavior, and gain operational insights. Another common use case is in **machine learning**, where data lakes serve as repositories for the diverse and rich datasets required to train AI models effectively. Additionally, they are instrumental in **Internet of Things (IoT)** ecosystems, where they collect and process sensor data from connected devices, enabling predictive maintenance, anomaly detection, and real-time monitoring.

### 5.4 Tools and Technologies Used

To support the efficient storage and processing of data, data lakes rely on a variety of tools and technologies, such as:

- **Storage Solutions:** AWS S3, Azure Data Lake Storage, Google Cloud Storage, Hadoop HDFS
- **Processing Engines:** Apache Spark, Apache Flink, Presto, Hive, Databricks
- **Data Management:** Apache Atlas (metadata management), Apache Ranger (security)
- **Analysis & Querying:** Trino (formerly Presto), Apache Drill, Athena (for querying data in S3)

When choosing between data lakes and data warehouses, organizations must consider specific needs, such as data types, performance requirements, and available budget. Data lakes are suitable for applications requiring flexibility and raw data storage, whereas data warehouses are ideal for structured analyses and rapid reporting. A hybrid approach, combining both solutions, can offer the flexibility and performance needed to maximize data processing capabilities. [3]

## 6 The Role of Analytical Models in the Context of Big Data

In a digital landscape defined by uncertainty, rapid change, and massive amounts of information, an organization's ability to understand what is happening in real-time and anticipate future events becomes a crucial differentiator. This is where analytical models come in, enabling the extraction of value from data through advanced and methodical analysis. These models can address fundamental questions about actions taken—why they happened, what caused them, what is likely to follow, and what actions should be taken next. Analytical models are traditionally grouped into four categories: **descriptive analytics**, **diagnostic analytics**, **predictive analytics**, **prescriptive analytics**. These types of analysis are often used in combination rather than isolation, as they are complementary and frequently integrated into the analytical workflows of modern Business

Intelligence platforms. In particular, predictive and prescriptive analytics have experienced accelerated development in the context of Big Data due to their ability to anticipate outcomes, behaviors, or trends based on historical data.

### 6.1 Descriptive Analytics

Descriptive analytics focuses on determining the outcome of an action—essentially answering the question “what happened?” This is a foundational form of analysis used to understand the past based on historical data. The goal is to provide a clear and concise overview of an organization's performance. [4]

**Techniques and tools** could be traditional BI dashboards and reports (Power BI, Tableau), data visualizations (line charts, pivot tables) or key performance indicators (KPIs). These types of analytics provide decision-makers with a retrospective view of performance, helping to identify patterns, seasonality, and results.

### 6.2 Diagnostic Analytics

Diagnostic analytics aims to explain why a particular outcome occurred. It enhances descriptive analysis by identifying the causes behind results and changes.

**Examples** could be sudden drops in sales in a specific region, increased customer churn rates or rising operational costs in a particular month.

**Techniques and tools** are usually drill-down analyses, correlation analyses and data segmentations.

### 6.3 Predictive Analytics

Predictive analytics focuses on determining what is likely to happen. It uses historical data, statistical models, and machine learning algorithms to forecast future trends.

**Examples** could be product demand forecasts, risk estimations for fraud or identification of customers who are likely to churn.

**Techniques and tools** are neural networks, regression, decision trees,



machine learning algorithms (scikit-learn, Azure ML), models trained on historical data and tested on new data. The purpose of predictive analytics is to help anticipate risks and opportunities, optimizing marketing, operational, and financial decisions.

The purpose of predictive analytics is to help anticipate risks and opportunities, optimizing marketing, operational, and financial decisions. In practice, these models can be integrated into business workflows to provide real-time insights, supporting proactive strategies rather than reactive ones.

#### 6.4 Prescriptive Analytics

Prescriptive analytics helps determine what action to take following an event or outcome. It combines predictive analysis with scenarios, business rules, and simulations to recommend concrete actions. [5]

**Examples** could be choosing the best marketing campaigns for a customer segment, optimizing delivery routes or budget simulation-based cost-cutting recommendations.

**Techniques and tools** are optimization algorithms (linear programming, Monte Carlo simulations), artificial intelligence or decision support systems and tools like IBM Decision Optimization, SAP Analytics Cloud. The role of this is to offer actionable recommendations, not just forecasts, enabling decisions based on modeled results.

Predictive analytics uses mathematical models, statistical algorithms, and machine learning techniques to identify patterns in data and generate predictions. These predictions can span various domains. Unlike descriptive analytics, which provides a retrospective view, predictive models rely on correlations and trends to offer forward-looking insights. Depending on the domain, predictive models can perform **classification** (e.g., estimating the likelihood of a customer churning) or

**regression** (e.g., forecasting sales for an upcoming period).

### 7 Algorithms Used in Predictive Models

The algorithms used in predictive models vary depending on the type of problem being addressed and the nature of the available data. Among the most widely applied is **linear regression**, which estimates continuous numeric outcomes based on the relationship between independent and dependent variables. **Decision trees** are also frequently employed, particularly in classification tasks, due to their interpretability and ease of use for non-technical stakeholders. For more complex data patterns, **artificial neural networks** offer high performance, especially in domains such as image recognition, natural language processing, and behavioral forecasting.

Another commonly used approach is **K-Means**, an unsupervised learning algorithm that segments data into distinct clusters based on similarity, proving useful in market segmentation and anomaly detection. More sophisticated ensemble methods such as **Random Forest** and **XGBoost** are favored for their robust predictive capabilities, handling high-dimensional data effectively and delivering strong performance across various industry applications. These algorithms are typically trained on historical data and validated on separate test sets to evaluate their generalization ability and ensure reliable deployment in real-world scenarios.

#### 7.1 Linear Regression

Linear regression is a supervised learning algorithm used to estimate a linear relationship between a dependent variable (target,  $Y$ ) and one or more independent variables (predictors,  $X_1, X_2, \dots, X_n$ ). It is considered a starting point in predictive modeling due to its simplicity and interpretability. [6]

**General equation:**

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n + \varepsilon$$

Where:

- **Y**: the dependent variable (the one being predicted)
- **X<sub>1</sub>...X<sub>n</sub>**: independent variables (predictors)
- **β<sub>0</sub>**: intercept (value of Y when all Xs are 0)
- **β<sub>1</sub>...β<sub>n</sub>**: regression coefficients (measure the impact of each predictor)
- **ε**: residual error

**Example:** a retail company wants to analyze the impact of its promotional budget on monthly sales. Using linear regression, it can build a model that predicts sales based on the budget, helping with more efficient marketing investment planning.

**Steps to apply linear regression:**

1. **Data collection** (e.g., sales, budget, season, region);
2. **Data exploration** (e.g., scatter plots, histograms);
3. **Model training** (estimating the β coefficients);
4. **Validation** (using R<sup>2</sup>, MSE, or RMSE);
5. **Result interpretation and implementation** in BI systems.

**Performance metrics:**

- **R<sup>2</sup> (Coefficient of Determination)** – shows how much of the variance in Y is explained by X;
- **MSE (Mean Squared Error)** – the average of squared differences between actual and predicted values;
- **MAE (Mean Absolute Error)** – the average of absolute differences between actual and predicted values.

**Advantages of linear regression** are high **interpretability** – easy to explain, **efficiency** – fast to train and strong **theoretical foundation**.

**Limitations** of this approach are that it usually requires a **linear relationship** between variables, it's **sensitive to outliers**, which can distort results, it assumes **independence among predictors** – multicollinearity can reduce validity and it cannot capture **complex interactions** between variables.

**Simple implementation in Python:**

```
from sklearn.linear_model import
LinearRegression
model = LinearRegression()
model.fit(X_train, Y_train)
pred = model.predict(X_test)
```

**In R:**

```
model <- lm(Sales ~ Budget, data = mydata)
summary(model)
```

**7.2 Decision Trees**

Decision trees are supervised learning algorithms used for both classification and regression tasks. They operate based on a set of logical rules (if-then-else) that split the data into subsets according to the values of specific attributes. The visual representation resembles an “inverted tree,” with the root representing the main condition, and the leaves corresponding to the final predictions. The algorithm evaluates all available features, selects the one that best splits the data (based on a splitting criterion), and creates a “branch” for each possible outcome. This recursive process continues for each data subset until certain stopping conditions are met: a maximum depth is reached, a node contains too few samples, or all samples in a node belong to the same class (for classification tasks). Common splitting criteria include **Gini Impurity** (measures how “impure” a node is), **Entropy & Information Gain** (used in algorithms like ID3 and C4.5), and **Mean Squared Error** (used for regression trees). [7]

**Practical example:**

A telecommunications company wants to determine whether a customer is at high risk of churning. A decision tree can be used to analyze features such as the number of calls to customer support,

contract duration, overdue bills, and satisfaction scores. The tree may learn rules like:

IF satisfactor\_score < 3 AND contract < 12 -> quitting

This logic can easily be integrated into a BI dashboard with automatic alerts.

**Advantages of decision trees** are that they're intuitive and have a visual structure, they do not require data scaling or standardization, they can handle both numerical and categorical data and they are able to perform well on small to medium-sized datasets.

**Limitations** are **overfitting** – trees tend to memorize the training data if not pruned, **instability** – small data changes can lead to significant structure changes, **lower performance** on large datasets compared to ensemble models.

#### Simplified implementation in Python:

```
from sklearn.tree import DecisionTreeClassifier
from sklearn.tree import plot_tree
model = DecisionTreeClassifier(max_depth=3)
model.fit(X_train, Y_train)
plot_tree(model, feature_names=features,
class_names=["NO", "YES"])
```

#### In R:

```
library(rpart)
model <- rpart(Risc ~ ., data = df, method =
"class")
plot(model)
text(model)
```

#### Performance metrics:

- **Accuracy** – percentage of correct predictions
- **Precision / Recall / F1-score** – especially useful in imbalanced classification problems
- **R<sup>2</sup> / MAE / RMSE** – when decision trees are used for regression

Example of a classification decision tree (Iris Dataset)

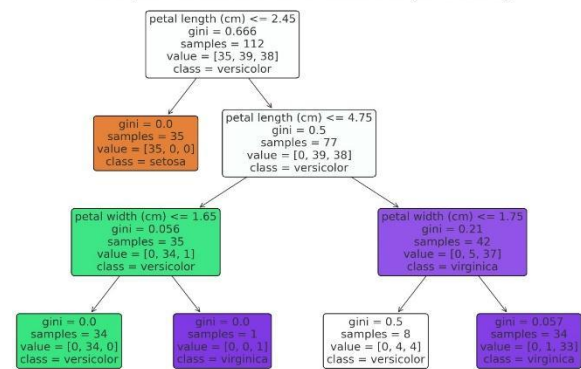


Fig 2. Decision Tree for the Iris Dataset

A classic example of decision tree structure is demonstrated using the **Iris dataset**, introduced by Ronald Fisher in 1936. It contains 150 records of three flower species and is a benchmark widely used to test classification algorithms. It is available in many popular libraries such as scikit-learn (Python) and R datasets.

### 7.3 Artificial Neural Networks (ANNs)

Artificial Neural Networks (ANNs) are machine learning models inspired by the structure and functioning of the human brain's neural network. They consist of artificial "neurons" organized into layers, which communicate with each other through mathematically simulated "synapses" (weights). ANNs are used for recognizing complex patterns and are highly effective in tasks such as classification, regression, image recognition, and natural language processing. [8]

A typical neural network includes: **input layer** – receives raw input data (e.g., numerical features), **hidden layers** – process the data through activations and weighted connections, **output layer** – generates the prediction (e.g., a class label or a numerical value).

Each neuron computes:

$$Z = w_1 * x_1 + w_2 * x_2 + \dots + w_n * x_n + b$$

$$a = f(z)$$

Where  $f$  is an **activation function** (e.g., ReLU, sigmoid, softmax).

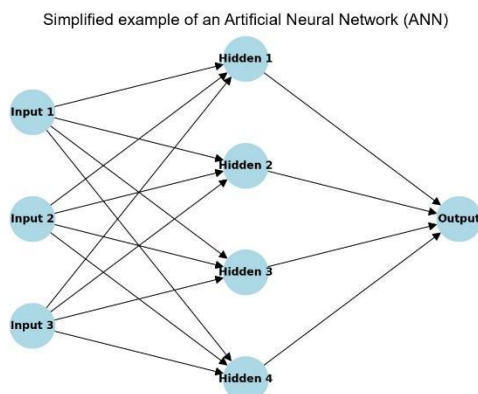
**Example:**

An e-commerce company may use a neural network to predict whether a customer will complete a purchase. The model might take into account factors such as time spent on the site, order history, product clicks, and device type. The network then outputs a probability between 0 and 1.

Common types of neural networks:

- **ANN (Artificial Neural Network)** – a simple, fully connected network
- **CNN (Convolutional Neural Network)** – used for image analysis
- **RNN (Recurrent Neural Network)** – used for sequential data (e.g., text or time-based transactions)

For business intelligence (BI), the most commonly used are ANNs and LSTMs (a type of RNN), often applied in forecasting and behavioral analysis.



**Fig 3.** Visual Example of a Neural Network

**Advantages** of using neural network algorithms are detection of complex and non-linear relationships, high scalability for large datasets, versatility (can be applied to classification, regression, text/image generation, etc.) and high effectiveness in behavioral prediction.

**Limitations** are **long training times** and **large datasets** for optimal performance,

demand of **significant computing power**, **overfitting** without proper regularization, **difficulty in interpretation**, which can hinder explainability in business settings.

### Simple Implementation in Python (using Keras and TensorFlow):

```
from keras.models import Sequential
from keras.layers import Dense
model = Sequential()
model.add(Dense(16, input_dim=4,
activation='relu'))
model.add(Dense(8, activation='relu'))
model.add(Dense(1, activation='sigmoid'))
model.compile(loss='binary_crossentropy',
optimizer='adam', metrics=['accuracy'])
model.fit(X_train,y_train,epochs=100,batch_size=10)
```

Performance Metrics for Neural Networks:

- **Accuracy, Precision, Recall, F1-score** – for classification tasks
- **Cross-Entropy Loss** – to evaluate learning progress
- **ROC Curve and AUC** – measure the model's ability to classify across all possible thresholds

Real-world applications of neural networks could be **Google Ads** – estimating click-through rates, **Netflix (and other online streaming services)** – personalized recommendations based on user viewing history, **banking** – classifying loan applications based on risk levels, **retail** – sales forecasting and facial recognition in stores.

### 7.4 K-Means – Clustering

K-Means is an unsupervised learning algorithm used to group data based on similarity. Unlike classification algorithms, K-Means does not rely on labels and instead, it automatically identifies natural groupings (clusters) in the data. [9]

**Steps of the algorithm:**

1. Select a number **k**, representing the desired number of clusters;

2. The algorithm randomly chooses **k** initial points called centroids;
3. Each data point is assigned to the nearest centroid using **Euclidean distance**;
4. Centroids are recalculated as the mean of the points in each cluster;
5. Steps 3–4 are repeated until the centroids no longer change significantly (convergence is reached).

#### Requirements and inputs:

- The number of clusters **k** must be chosen before running the algorithm;
- Works best with **numerical, scaled data**;
- Assumes **spherical clusters of similar size**, and is not sensitive to data distribution.

**Advantages** of this algorithm: swift implementation and adaptability for large datasets, efficiency and interpretability when data has clear structure, huge usage in **segmentation tasks** (e.g., market or customer segmentation).

**Limitations** could be choosing the value of **k**, which is subjective (can be aided by the **Elbow Method** to ensure minimal acceptable efficiency), the **sensitivity to outliers** and **random initialization**, and incapability for **categorical data** (unless preprocessed via numerical encoding).

#### Performance metrics:

- **Silhouette Score** – measures how well each point fits within its cluster;
- **Inertia (Sum of Squared Errors – SSE)** – total distance between points and their cluster centroids;
- **Inter-cluster distances** – measure the separation between clusters.

#### Simple implementation in Python:

```
from sklearn.cluster import KMeans
```

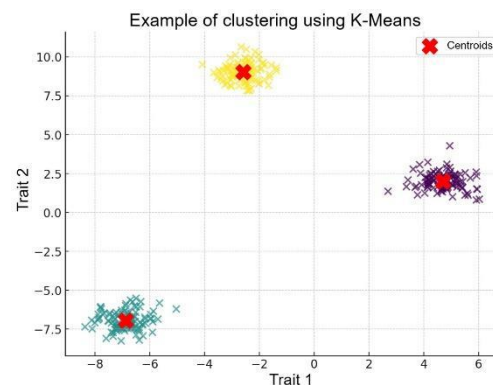
```
model = KMeans(n_clusters=3)
model.fit(X)
labels = model.predict(X)

import matplotlib.pyplot as plt
plt.scatter(X[:, 0], X[:, 1], c=labels,
            cmap='viridis')
```

#### In R:

```
kmeans_result <- kmeans(mydata, centers = 3)
plot(mydata, col = kmeans_result$cluster)
```

In **Power BI**, K-Means clustering can be performed by integrating with **Python scripts**, or externally calculated cluster labels can be imported and used for visualizations and analysis.



**Fig 4.** Visual Example (Chart) of Clustering

## 8 Modern Tools Supporting Predictive Analytics

As data analytics becomes increasingly accessible, a growing number of Business Intelligence (BI) platforms have begun to incorporate integrated predictive capabilities. Among the most prominent solutions is **Power BI** by Microsoft, which allows users to embed predictive models developed in **Azure Machine Learning** or execute custom Python scripts directly within the platform. Similarly, **Tableau** offers support for integrating external predictive models into dashboards and enables connectivity with programming environments such as Python—through the **TabPy** server—or R, allowing for advanced statistical and machine learning functions. **IBM Watson Analytics** distinguishes itself by providing automated

analysis features that assist users in identifying meaningful patterns within their data without requiring deep technical expertise. Another widely used platform is **RapidMiner**, a comprehensive data science environment designed to be user-friendly, even for individuals without a background in programming. It supports a broad range of predictive modeling tasks through its visual workflow interface.

Also, **Qlik Sense** emphasizes exploratory data visualization while also supporting advanced analytics through extensibility features, which enable the integration of predictive models into interactive reports and applications. Collectively, these tools empower organizations to construct tailored predictive models, embed them seamlessly into operational workflows, and generate actionable insights in real time—enhancing decision-making capabilities across various business functions.

### 8.1 Challenges in Applying Predictive Models

Despite their growing importance and proven effectiveness in many domains, predictive models present several notable challenges that organizations must address. One of the most critical issues is **data quality**, as accurate and reliable predictions rely heavily on datasets that are clean, relevant, and complete. In the absence of such data, the models' outputs may be misleading or invalid.

Another significant challenge lies in the **technical complexity** of many predictive algorithms. Certain models—especially those based on deep learning—are often perceived as “black boxes” due to their opaque internal logic. This lack of transparency can hinder user trust and reduce their applicability in contexts where interpretability is essential.

Additionally, predictive models are prone to **overfitting**, a phenomenon where the model becomes overly tailored to the training data, capturing noise rather than general patterns. As a result, performance

deteriorates when the model is applied to new, unseen data.

The deployment of advanced predictive models also requires considerable **infrastructure and financial investment**, including robust hardware, scalable storage, and specialized software tools. These requirements may limit adoption, particularly for smaller organizations.

The increased use of automated decision-making introduces **ethical and transparency concerns**, particularly in sensitive domains such as credit scoring, hiring, and law enforcement. These systems can inadvertently reinforce existing biases or produce discriminatory outcomes, making accountability and explainability critical aspects of model governance.

## 9 Benefits of Using Big Data and Business Intelligence

BI and Big Data allow organizations to make decisions based on solid evidence, thus reducing risks associated with uncertainty. By leveraging predictive analytics and machine learning algorithms, companies can forecast market trends and customer behavior.

### 9.1 Personalizing Customer Experience

By collecting and analyzing data on customer preferences and behavior, companies can deliver personalized products and services. This leads to increased customer satisfaction and loyalty.

### 9.2 Operational Optimization

Big Data helps identify inefficiencies in supply chains, production, and logistics. Data analysis supports cost reduction, improved productivity, and resource optimization.

### 9.3 Fraud Detection and Enhanced Security

By analyzing transactional and behavioral data, organizations can detect suspicious patterns that may indicate fraud or



cyberattacks, helping to improve security and protect sensitive information.

#### 9.4 Improving Marketing Efficiency

BI enables market segmentation and the development of targeted marketing campaigns, increasing conversion rates and maximizing return on promotional investments.

### 10 Challenges of Using Big Data and Business Intelligence

One of the biggest obstacles to adopting Big Data is the enormous volume of information generated daily. Organizations must invest in appropriate infrastructure and advanced technologies for data storage and processing.

#### 10.1 Data Quality

Collected data may contain errors, duplicates, or incomplete information, which can lead to faulty decisions. Implementing effective data cleaning and validation strategies is essential.

#### 10.2 Security and Regulatory Compliance

Organizations must comply with data protection regulations such as GDPR, which involves adopting strict security and data anonymization policies.

#### 10.3 Implementation Complexity

Integrating Big Data and BI solutions requires technical expertise, robust infrastructure, and a shift in organizational culture to fully harness the potential of these technologies.

#### 10.4 Resistance to Change

Employees may be reluctant to adopt new technologies. An organizational change strategy, including training and awareness sessions, is necessary to ensure effective adoption of BI and Big Data tools.

#### 10.5 The Synergy of Big Data and Business Intelligence for Competitive Advantage



**Fig 5.** Representative Chart of Business Intelligence Solutions

The integration of Big Data and Business Intelligence (BI) represents a strategic asset for organizations seeking to maintain a competitive position in dynamic markets. By combining the vast processing capabilities of Big Data with the structured analytical frameworks of BI, companies are able to implement advanced analytical techniques, including machine learning and artificial intelligence, to extract deeper insights from complex datasets. This integration facilitates the automation of decision-making processes, thereby enabling faster and more informed responses to rapid shifts in market conditions [10]. Furthermore, the ability to derive actionable insights from real-time data allows organizations to better understand and anticipate customer demands, ensuring improved alignment between business strategies and consumer expectations [11]. Ultimately, this synergy contributes to increased organizational agility, empowering firms to adapt swiftly to emerging challenges and capitalize on new opportunities as they arise [12].

From a financial perspective, the impact is equally remarkable. Companies using Big Data report significant reductions in storage and processing costs, as well as noticeable improvements in return on investment (ROI). Through advanced data



analysis, organizations can identify new business opportunities, optimize marketing strategies, and increase customer conversion rates. For example, in retail, Big Data enables not only product recommendation personalization but also more efficient inventory management, reducing waste and maximizing profit.

The competitive advantages are particularly valuable. Big Data enables organizations to develop innovative products and services based on a deep understanding of consumer needs and behavior. Moreover, with real-time data analysis, companies can respond quickly to market changes, adjusting strategies to stay relevant. Data-driven marketing also allows for more precise and efficient campaigns, while a holistic organizational view helps remove departmental silos, boosting collaboration and efficiency.

However, implementing Big Data is not without challenges. One of the greatest difficulties is managing the large volumes of data, which require powerful hardware infrastructures and sophisticated software solutions. Data quality is another critical aspect—missing or inaccurate information can lead to false conclusions and poor decisions. Integrating data from various sources, such as ERP systems, CRMs, or social networks, is also complex and demands specialized tools and technical expertise.

Data security and privacy are major concerns as well. As organizations collect and store more information, they become more vulnerable to security breaches. Compliance with regulations such as GDPR is essential to avoid penalties and maintain customer trust. Additionally, the shortage of skilled Big Data professionals remains a persistent problem, requiring companies to invest in training and professional development to build competent teams. [13]

## 11. Tools for Managing Big Data

The tools and platforms used for managing Big Data play an essential role in

addressing the technical and operational challenges associated with volume, velocity, and variety. One of the foundational technologies in this ecosystem is **Hadoop MapReduce**, a distributed processing framework that enables the execution of large-scale data processing tasks across clusters of commodity hardware. It operates on a two-phase model—Map and Reduce—that supports parallel computation and fault tolerance, making it suitable for batch processing of unstructured and semi-structured data. [14]

Building on some of the limitations of Hadoop, **Apache Spark** emerged as a high-performance in-memory computing engine capable of both batch and real-time data processing. Spark supports a variety of APIs, including SQL, Python, Scala, and R, and includes modules for streaming, machine learning, and graph processing. Its ability to process data in-memory drastically reduces latency, making it well-suited for real-time analytics and iterative algorithms commonly used in data science and BI contexts. [15]

In the area of data visualization, **Tableau** has become one of the most widely adopted tools. It enables users to create interactive dashboards and intuitive visual representations of data, even without extensive programming knowledge. Tableau's real-time data connectivity and drag-and-drop interface make it a preferred choice for business analysts aiming to uncover trends and patterns quickly and present them to stakeholders. [16]

For data storage, platforms such as **HBase** and **Apache Cassandra** offer distributed and scalable solutions tailored for Big Data environments. HBase, built on top of Hadoop's HDFS, is optimized for random read/write access to large datasets and is particularly effective in environments requiring strong consistency guarantees [17]. In contrast, Cassandra is designed for high availability and horizontal scalability across multiple data centers. Its

decentralized architecture and support for eventual consistency make it ideal for applications that prioritize uptime and speed over strict consistency. [18]

When dealing with data relationships that go beyond tabular formats, **Neo4j** offers a graph-based storage paradigm that excels at modeling and analyzing complex networks. It is especially useful in use cases such as fraud detection, recommendation engines, and supply chain optimization, where the connections between entities are as important as the entities themselves. [19]

Technologies such as **JSON** and **NoSQL databases** provide flexible schemas and are widely used for storing semi-structured or unstructured data. These tools allow for the dynamic inclusion of new fields and formats, facilitating the integration of heterogeneous data sources and supporting agility in data modeling. [20] Complementing these, **RESTful APIs** have become the standard for application integration, enabling seamless communication between Big Data platforms, web services, and BI applications. They support system interoperability and modular architecture design by providing standardized, stateless communication protocols.

Looking ahead, the Big Data landscape is expected to be further transformed by ongoing advances in artificial intelligence (AI), machine learning (ML), and the Internet of Things (IoT). These technologies are contributing to increasingly autonomous and adaptive systems capable of deriving insights in real time. Additionally, **quantum computing**, although still in its infancy, holds the potential to revolutionize data processing by solving highly complex problems at speeds that are unattainable with classical architectures [21]. Organizations that embrace and integrate these emerging technologies into their data strategy are likely to achieve significant and sustainable competitive advantages, while

those that fail to adapt may face increasing difficulty in remaining relevant.

## 12. Conclusions

The effective implementation of Big Data and BI is a critical success factor for modern organizations. Although there are significant challenges, the benefits far outweigh these obstacles—enabling companies to make better decisions, understand customer behavior more deeply, and optimize operations. The synergy between Big Data and BI is key to achieving sustainable competitive advantage in a dynamic and complex business environment. The importance of data in decision-making could be summarized to be the proportion of decisions made based on data versus those based on intuition. According to Gartner's 2019 *Magic Quadrant* for Business Intelligence and Analytics Platforms, the leading providers in this space are **Tableau**, **Microsoft**, and **Qlik**.

## References

- [1] Alnoukari, M. (2020). *From Business Intelligence to Big Data: The Power of Analytics*. DOI: 10.4018/978-1-7998-5781-5.ch003
- [2] Paul, C., Shama, V., & Laisis, R. (2022). *ETL in the Era of Big Data: Challenges and Solutions*
- [3] John, B. (2025). *Data Processing in Data Lakes vs. Data Warehouses*
- [4] Sharda, R., Delen, D., & Turban, E. (2020). *Analytics, Data Science, & Artificial Intelligence*
- [5] Taylor, J. (2019). *Decision Management Systems: A Practical Guide to Using Business Rules and Predictive Analytics*. IBM Press
- [6] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning*. Springer
- [7] Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques*. Elsevier

- [8] Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly
- [9] Tan, P.-N., Steinbach, M., & Kumar, V. (2019). *Introduction to Data Mining*. Pearson
- [10] SHIFT ASIA. (n.d.). *Unlocking the Synergy of Big Data and Business Intelligence*.
- [11] Decision Foundry (n.d.). *Big Data and Business Intelligence: Unveiling the Impact*.
- [12] Al-Darras & Tanova (2022). *From Big Data Analytics to Organizational Agility: What Is the Mechanism?*
- [13] Jeren Agh (2020) TechGDPR - *The Impact of the GDPR on Big Data*. Accessed 28 June 2025: <https://techgdpr.com/blog/impact-of-gdpr-on-big-data/>
- [14] O'Reilly Media. (2015). *Hadoop: The Definitive Guide*
- [15] Databricks. (2023). *What is Apache Spark?*. Accessed 28 June 2025: <https://www.databricks.com/glossary/what-is-apache-spark>
- Jay Ripton. (2025). *TechRadar - Best data visualization tools of 2025*. Accessed 28 June 2025: <https://www.techradar.com/best/best-data-visualization-tools>
- [16] Apache HBase. (2022). *Use Cases & Architecture Overview*. Accessed 28 June 2025: <https://hbase.apache.org/>
- [17] Apache Cassandra. (2023). *Features & Use Cases*. Accessed 28 June 2025: [https://cassandra.apache.org/\\_/index.html](https://cassandra.apache.org/_/index.html)
- [18] Neo4j. (2022). *What is a Graph Database?*. Accessed 28 June 2025: <https://neo4j.com/docs/getting-started/graph-database/>
- [19] MongoDB. (2023). *Why NoSQL?*. Accessed 28 June 2025: <https://www.mongodb.com/resources/basics/databases/nosql-explained>
- [20] Josh Schneider, Ian Smalley, IBM Quantum. (2023). *What is quantum computing?*. Accessed June 28 2025: <https://www.ibm.com/think/topics/quantum-computing>



business operations

**Alexandru-Dan TURIAC** graduated the Faculty of Economic Cybernetics, Statistics and Informatics in 2023, and is currently in the final year of the master's program in Economic Informatics at the Bucharest University of Economic Studies, after completing a master's program in Databases for Business Support. Professionally, he works as a DevOps engineer, focusing on the automation, deployment, and monitoring of scalable software systems. His work involves ensuring system reliability, optimizing CI/CD pipelines, and integrating data-driven tools to support



particular focus on regulation and automation

**Mihai-Paul UNGUREANU** is currently a master student in the Databases for Business Support program at the Bucharest University of Economic Studies and is expected to graduate in 2025. He previously earned a bachelor's degree in 2023 from the Faculty of Economic Cybernetics, Statistics and Informatics, where he studied the Economic Informatics program. He has gained experience in web development, object-oriented programming and database technologies (SQL, NoSQL). His current interests include artificial intelligence algorithms and workflows, with a



**Mihnea-Cristian TACHE** graduated from the Faculty of Economic Cybernetics, Statistics and Informatics in 2023, and later will have completed a master's program in Database for Business Support at the Bucharest University of Economic Studies, in 2025. Currently in the last year as a master student in Economic-Informatics. Professionally, he works as a Frontend engineer, with responsibilities including the development and monitoring of scalable software systems.

## The Role of Data Science and Machine Learning in Strategic Decision Making

Elena-Luiza JALEA, Marian-Robert VLĂDUȚĂ  
Bucharest University of Economic Studies  
Faculty of Cybernetics, Statistics and Economic Informatics  
Bucharest, Romania  
[jaleaelena20@stud.ase.ro](mailto:jaleaelena20@stud.ase.ro), [vladutamarian20@stud.ase.ro](mailto:vladutamarian20@stud.ase.ro)

*Digital transformation accelerates decision-making processes in modern organizations, generating a massive volume of data that must be processed quickly and efficiently. In this context, Data Science and Machine Learning are no longer merely technical tools but are becoming fundamental components of business strategies.*

*The application of advanced data analysis technologies enables the identification of patterns, trends, and correlations that can significantly influence strategic decision-making. However, integrating these technologies into organizational processes also raises a series of challenges related to interpretability, ethics, accuracy, and adaptability.*

*This paper analyzes, from both a theoretical and practical perspective, how Data Science and Machine Learning contribute to the improvement of the decision-making process in organizations, while also addressing their limitations in a constantly changing economic environment.*

**Keywords:** Data Science, Machine Learning, strategic decisions, predictive analysis, digital transformation, Business Intelligence.

### 1 Introduction

In the era of accelerated digital transformation, data has become one of the most valuable resources of modern organizations. The huge volumes of information generated from online interactions, internal systems, or external sources put pressure on traditional analysis and decision-making processes. In this context, the ability to extract value from data is no longer an optional competitive advantage, but a strategic necessity [1]. Data Science is emerging as an essential interdisciplinary field that combines statistical methods, computer science, and analytical thinking to identify patterns, relationships, and relevant insights in data. In parallel, Machine Learning, a key component of artificial intelligence, enables algorithms to learn from data and make automatic predictions or classifications without being explicitly programmed for each situation [2]. Thus, these two fields offer valuable tools for strategic decision-making in

organizations, from supply chain optimization to consumer behavior analysis, risk management, or intelligent resource allocation. However, the integration of these technologies also raises a series of challenges: lack of decision-making transparency, the risk of overreliance on algorithmic models, ethical issues, or lack of internal expertise [3].

This paper aims to explore the role of Data Science and Machine Learning in supporting strategic decisions, emphasizing their transformational potential as well as the limitations that may affect the effectiveness of their application in real organizational contexts.

### 2. The Conceptual Foundations of Data Science and Machine Learning-Assisted Decisions

Before analyzing the actual impact of these fields on the strategic decision-making process, a clear understanding of the fundamental concepts is necessary.

Data Science is an interdisciplinary field that integrates statistical methods, machine learning algorithms, data processing procedures, and interactive visualization, with the goal of extracting useful knowledge from large volumes of data. This field brings together expertise from mathematics, computer science, data science, and domain knowledge, being oriented towards solving real-world problems based on data analysis [4].

An essential element of Data Science is the analytical pipeline process, which involves several stages: data collection and cleaning, data exploration and understanding, predictive modeling, and model performance evaluation. In practice, Data Science aims to transform raw data into actionable information for decision-making.

Within this ecosystem, Machine Learning (ML) represents a specialized subset that focuses on developing algorithms capable of "learning" from historical data to make automatic predictions or classifications. ML is based on statistical and computational principles and includes techniques such as linear regression, decision trees, neural networks, or ensemble algorithms [5].

Unlike classical programming, where the logic is explicitly written by the developer, in Machine Learning the model learns automatically based on examples and generalizes new behaviors. This ability makes ML extremely valuable in contexts such as sales forecasting, risk analysis, or service personalization [6].

Although different in approach, Data Science and Machine Learning are complementary and work together to support modern decision-making processes, especially in organizations focused on innovation and strategic adaptability.

Currently, the applications of Data Science and Machine Learning go beyond the technical sphere and penetrate all organizational levels. From optimizing supply chains and analyzing customer

behavior to supporting strategic management decisions, these technologies are increasingly becoming an integral part of companies' operational infrastructure. This is reflected not only in technological investments but also in skill requirements: many organizations already require decision-makers to have a basic understanding of data modeling, evaluation, and interpretation concepts.

Moreover, the increasing accessibility of open-source tools (such as Python, R, Jupyter Notebooks, scikit-learn, or TensorFlow) is democratizing the analytical process and paving the way toward a data-driven organizational culture. This culture involves not only using data to support decisions but also developing an analytical mindset within teams, where data is critically interpreted, models are validated, and intuition is complemented by statistical rigor.

Thus, the theoretical understanding of Data Science and Machine Learning is not merely an academic component, but a practical necessity for professionals involved in decision-making. These disciplines form the conceptual and operational framework of the new decision-making paradigms, based on knowledge, automation, and adaptability.

### **3. Decision-Making Models and the Evolution from Business Intelligence to Data Science**

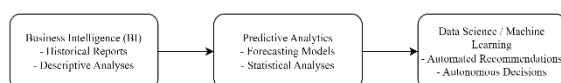
The decision-making process within organizations has undergone significant transformation over the past two decades, alongside the evolution of technologies for data collection, storage, and analysis. In the past, strategic decisions were based on managers' experience, monthly accounting reports, and empirical observations. Although sometimes effective, this approach was often subjective and reactive, lacking the ability to anticipate future developments.

With the emergence of decision support information systems, the concept of Business Intelligence (BI) was born—a set

of tools and methodologies that enables the organization, analysis, and visualization of historical data in a way that is accessible to decision-makers. BI primarily provides descriptive analysis: what happened, where variations occurred, how key indicators performed [7].

However, Business Intelligence remains limited to what can be described retrospectively. In a dynamic business environment, companies can no longer wait for monthly reports to react. This is where the natural evolution toward Data Science comes into play, adding predictive and prescriptive capabilities. By integrating Machine Learning techniques, advanced statistical models, and automated knowledge extraction processes from data, organizations can anticipate behaviors, identify risks in advance, or propose optimized actions [8].

The major difference between BI and DS is the level of autonomy and added value in the decision-making process. While BI requires human interpretation of reports, DS provides automated recommendations, quantifiable predictions, and continuous learning. Moreover, Data Science is not limited to structured data (e.g., databases) but also integrates unstructured data (text, images, sound), allowing for a much deeper contextual understanding. This is clearly illustrated in the figure below:



**Fig. 1.** The Transition from *BI* to *DS* in Decision-Making

In the figure above, we can observe the natural transition from descriptive systems to those that not only explain the past but also model the future. At the same time, Data Science enables decision personalization at the level of the customer, product, or moment, increasing efficiency and relevance.

Another important aspect is the real-time integration of data, made possible by Big Data and Cloud infrastructure, which transforms decisions from periodic processes into continuous flows, assisted by algorithms that constantly learn and adapt [9].

This transformation does not completely eliminate the human factor but elevates it to a higher level: from manual analysis to the validation of automatically generated proposals. Thus, Data Science not only supports decision-making but transforms it into a systematic, repeatable, and scalable process.

In this context, it is important to understand not only the benefits brought by these new approaches but also how they differ conceptually and functionally from traditional methods. Although many organizations use the terms Business Intelligence and Data Science interchangeably, in reality, there are significant differences in purpose, methodology, and decision-making impact.

For a better understanding of the essential differences between the two paradigms, Business Intelligence and Data Science, a synthetic comparison of their main characteristics is useful:

**Table 1.** Comparison Between Business Intelligence and Data Science in the Decision-Making Process

Characteristic	Business Intelligence (BI)	Data Science (DS)
Type of analysis	Descriptive	Predictive and prescriptive
Data used	Structured, historical	Structured and unstructured
Decision frequency	Periodic (e.g., monthly)	Continuous / real-time
Level of automation	Low – requires human interpretation	High – automatic capabilities
Output	Reports, dashboards	Predictive models, recommendations, simulations
Data sources	Internal	Internal + external



The table highlights the fact that although BI and DS can coexist within the same organization, they serve different purposes and address distinct needs. Business Intelligence provides clarity about the past, while Data Science opens up perspectives on the future. The integration of both approaches can bring a significant competitive advantage to organizations that aspire to make intelligent, fast, and data-driven decisions.

Moreover, unlike Business Intelligence, which provides punctual and static decision support, Data Science enables the construction of adaptive decision systems capable of automatically adjusting as new data emerges or contextual parameters change. This flexibility is crucial in dynamic industries such as e-commerce, financial services, or logistics, where reaction times must be minimal and the level of personalization, maximal.

Last but not least, Data Science brings about a fundamental mindset shift within organizations. While BI encourages reactive thinking based on past results, ML- and AI-based approaches promote a proactive decision-making culture centered on anticipation, continuous testing, and iterative optimization. In this way, data is no longer just a tool for justifying decisions, but becomes an essential element in building future strategies.

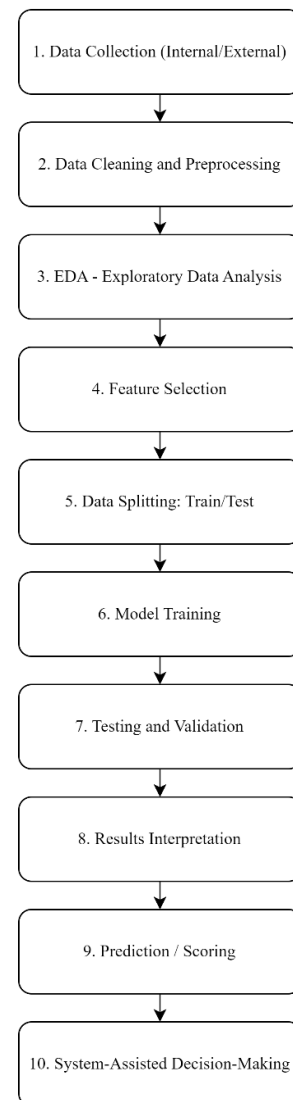
#### 4. The Role of Machine Learning in Strategic Decision-Making

In an increasingly complex organizational environment, the ability to make fast, well-founded, and adaptable decisions becomes essential. In this context, Machine Learning (ML) plays a central role by enabling information systems to automatically learn from data, identify patterns, and make predictions with a high degree of accuracy. This offers much more dynamic and efficient decision support compared to traditional methods. ML is used in strategic decision-making across multiple areas of activity, such as customer behavior analysis, fraud detection, sales

forecasting, supply chain optimization, or financial risk assessment. Models can be trained to recognize complex scenarios, understand multiple variables, and deliver results in real time—key elements in a modern decision-making ecosystem.

One of the most important advantages of ML is its ability to generalize. Unlike fixed rules, machine learning models can adapt to new, unfamiliar data, making them useful in volatile or unforeseen contexts. Additionally, models can be constantly recalibrated as new data accumulates, keeping decisions relevant and up to date.

To understand how this type of decision support works in practice, we present below a general workflow for an ML-assisted decision system.



**Fig. 2.** Machine Learning-Assisted Decision Workflow

This model enables evidence-based, automated, and scalable decision-making, significantly reducing the time required for manual analysis. Moreover, integrating these models into dashboard platforms or ERP systems allows results to be accessible to managers without advanced technical knowledge, thus democratizing the decision-making process.

However, the success of applying ML in decision-making depends on several factors: data quality, model interpretability, integration with organizational processes, and, last but not least, user trust in automated systems.

### 5. Advantages and Potential in Modern Organizations

The integration of technologies based on Data Science and Machine Learning into decision-making processes brings a series of significant advantages, both from an operational and strategic perspective. According to a recent study conducted by Venkatapathi and Vishnuvardhan (2025), organizations that actively implement these technologies have reported considerable increases in efficiency, prediction accuracy, and response speed in decision-making [10]. For example, the adoption of these technologies has led to:

- up to a 30% increase in sales forecast accuracy;
- a 25% reduction in losses caused by fraud in financial systems;
- acceleration of the decision-making cycle in organizations with advanced digital infrastructure.

These benefits are noticeable in the increased efficiency of processes, the ability to personalize offers, and the reduction of response time to environmental changes.

A major advantage is the automation of recurring decisions. ML models can be trained on historical data to recognize patterns and make decisions without human intervention, significantly reducing the effort required for repetitive activities

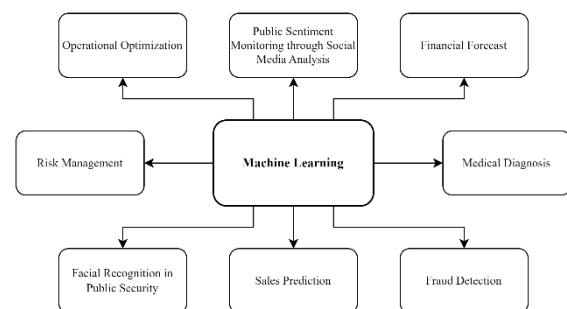
(e.g., automatic credit approval, fraud detection, or product recommendation).

Secondly, ML-based systems allow for high scalability: once developed, a model can be applied simultaneously across multiple areas of activity without requiring major adjustments. As a result, marginal costs decrease, and the benefits can extend across the entire organization.

Additionally, Data Science enables advanced decision personalization by integrating unstructured data such as website behavior, reviews, social media data, or even audio feedback. This level of granularity leads to a deeper understanding of customers and to decisions with greater impact in customer relations.

Another essential aspect is the ability of these technologies to identify opportunities and risks that are not obvious through classical analysis. For example, clustering models can detect new customer segments, and anomaly detection algorithms can flag subtle deviations from normal behaviors in financial or operational data.

In the figure below, some of the most important areas where Machine Learning brings strategic value to organizations are summarized:



**Fig. 3.** Strategic Application Areas of ML in Organizations

Last but not least, the use of these tools contributes to the development of a data-driven organizational culture, where intuition is complemented by evidence and decision-making becomes an iterative process supported by technology. This fosters innovation, transparency, and adaptability—essential qualities in a

constantly changing economic environment.

## 6. Limits, Risks, and Ethics of Using ML in Decision-Making

Although Machine Learning offers multiple benefits in strategic decision-making, the use of these technologies raises several important issues related to technical limitations, operational risks, and ethical dilemmas. A critical understanding of these aspects is essential for responsible and effective implementation.

One of the most discussed limitations is the dependence on high-quality data. ML models are heavily influenced by the datasets on which they are trained. Incomplete, biased, or irrelevant data can lead to incorrect or even dangerous decisions. For example, in the case of automated recruitment systems, a model trained on biased historical data may perpetuate gender or ethnic discrimination without users being aware [11].

Another critical aspect is the lack of transparency (a phenomenon known as “black-box models”). In many cases, complex models such as neural networks cannot intuitively explain *why* a certain prediction was made. This raises serious concerns in sensitive fields such as medicine or justice, where trust and accountability are essential.

There are also risks associated with the automation of decisions without human intervention, which can lead to a loss of control over the decision-making process or the amplification of errors. It is important that models be used as support tools, not as replacements for human reasoning, especially in situations with a high degree of uncertainty or ambiguity.

In addition, as ML is increasingly deployed on a large scale, there is an urgent need for algorithmic governance frameworks that establish clear rules regarding decision-making responsibility. When an erroneous decision made by an algorithm leads to negative consequences—for example, a denied loan

or a misdiagnosed medical condition—it becomes difficult to determine who is accountable: the algorithm developer, the organization using it, or the model itself? Therefore, organizations must implement policies for periodic model auditing, ensure transparency, and establish internal ethics committees capable of evaluating and regulating the impact of these technologies on the human factors involved.

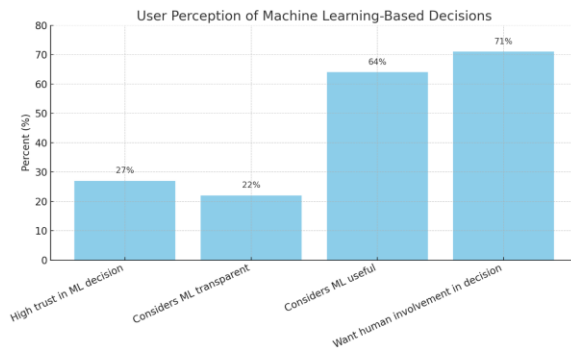
From an ethical standpoint, the discussion becomes even more complex. Questions arise regarding data privacy, user consent, the accountability of automated decisions, and even the right to explanation (a concept promoted by European AI and GDPR regulations).

In the following figure, the main risks and limitations associated with the use of Machine Learning in decision-making processes are summarized:

**Table 2.** Main Risks and Challenges in Using ML for Strategic Decision-Making

Category	Consequence
Biased data	Models perpetuate inequalities; discriminatory or incorrect decisions
Lack of transparency (black-box)	Inability to justify decisions; loss of trust from stakeholders
Low data quality	Erroneous predictions; poor model performance in real applications
Lack of model updating	Models become irrelevant over time; decisions based on outdated patterns
Excessive automation	Lack of human control; risk of propagating systemic errors
Ethical and legal issues	Violation of privacy and GDPR; lawsuits or legal sanctions

Beyond the risks identified at a theoretical level, real-world experience within organizations reveals a clear gap between the perceived usefulness and the actual trust in ML-assisted decisions. According to recent studies (PwC, IEEE), most organizations acknowledge the potential of these technologies, yet hesitate to fully transfer critical decisions to algorithms.



**Fig. 4.** User Perception of Transparency and Trust in ML-Assisted Decisions [12]

This data highlights the importance of building a solid ethical and operational framework in which ML models are continuously validated, explainable, and integrated into decision-making processes where the human factor has the final say. Without this balance, the use of algorithms may generate more uncertainty than clarity.

## 7. Case Studies and Examples

To understand the concrete impact of Data Science and Machine Learning technologies on the decision-making process, it is useful to analyze a few real case studies and applied examples in various organizational contexts. These highlight how machine learning models can bring added value to decision-making, as well as the practical challenges encountered during implementation.

### 1. Automated Credit Scoring – JPMorgan Chase (Financial Sector)

JPMorgan Chase is one of the largest financial institutions in the world and a leader in adopting artificial intelligence solutions. To streamline the credit risk assessment process, the bank developed models based on Random Forest and Gradient Boosting algorithms, which analyze hundreds of variables to estimate the probability of default.

The model largely replaced traditional static scoring systems and proved to be much more flexible and accurate in the volatile post-pandemic economic context. The system enables real-time evaluation of

credit applications, with reduced response time and increased adaptability to new customer behaviors.

Following the implementation of these models, the bank reported a 20% reduction in non-performing loans and a 12% increase in the number of eligible customers, indicating a direct impact on profitability and operational efficiency [13].

### 2. AI-Assisted Medical Diagnosis – IBM Watson Health (Medical Sector)

The IBM Watson Health system was designed to support physicians in making complex clinical decisions. Watson analyzes millions of scientific articles, the patient's medical history, clinical data, and test results to recommend personalized treatment options, especially in oncology. In a pilot study conducted in the United States, Watson identified alternative treatments in 30% of advanced cancer cases, offering options that had not previously been considered by medical teams. This type of assistance can be vital in cases where quick and well-informed decisions save lives.

However, the technology has also faced challenges, such as difficulties integrating into existing clinical workflows, the need for standardized data, and trust issues from medical staff. Even so, Watson remains a benchmark in ML-based decision support in the medical field [14].

### 3. Personalized Recommendations – Amazon (E-commerce Sector)

Amazon uses some of the most sophisticated Machine Learning-based recommendation systems in the world. The algorithms analyze behavioral data in real time: search history, purchases, ratings, time spent on a page, abandoned carts, etc. By using collaborative filtering methods, neural networks, and reinforcement learning, Amazon can accurately predict which product is most relevant for each user.

The system is responsible for approximately 35% of the sales generated on the platform, demonstrating a direct impact on strategic business decisions: inventory, marketing campaigns, and interface design [15].

This application shows how ML can transform a mundane decision-making process into an autonomous, adaptive, and highly profitable system.

#### 4. *Facial Recognition – Metropolitan*

##### *Police London (Public Security Sector)*

The Metropolitan Police in London has tested and implemented a real-time facial recognition (RTFR) system to detect individuals wanted by authorities in public spaces. Smart cameras are connected to a database containing photographs of suspects, and the system automatically flags matches.

In an official test conducted in 2020, the system correctly identified 8 out of 10 individuals in the database. However, the implementation was criticized by human rights organizations, which pointed out potential issues of racial bias and lack of public consent.

This application raises important questions about the balance between efficiency and civil liberties, being a clear example of the ethical challenges associated with the use of ML in sensitive public spaces [16].

#### 5. *Netflix – Personalized*

##### *Recommendation System*

Netflix has become emblematic for the use of Machine Learning in personalizing multimedia content. Its hybrid recommendation system combines collaborative filtering, clustering, NLP (for movie descriptions), and deep learning to anticipate each user's preferences.

The academic paper by Gomez-Uribe & Hunt (2015) details how Netflix's algorithms continuously evolve, learning from user interactions to decide not only what content to recommend but also how to visually present it. The result? Viewing

time increases, churn decreases, and the value per user rises consistently [17].

It is one of the best commercial examples of a strategic decision fully guided by ML in real time.

#### 6. *Biometric Surveillance – China (ML Applications in Public Security)*

In China, facial recognition and behavioral analysis technologies are used on an unprecedented scale. Millions of cameras connected to ML systems monitor traffic, citizen movements, and public interactions in cities such as Beijing, Hangzhou, and Shenzhen.

These systems are integrated into social scoring projects, where citizens' behavior influences access to services, credit, or even travel freedoms. Although the systems have significantly reduced certain types of crime, they have been criticized by the international community for lack of transparency, excessive surveillance, and the risk of abuse [18].

It is an extremely influential example of how Machine Learning can change the decision-making architecture at the state level, not just within companies.

These examples highlight that ML is already integrated into high-impact decisions, from finance and medicine to security and entertainment. At the same time, they emphasize the need for a balance between efficiency, ethics, and regulation, especially when decisions directly affect individuals or entire societies.

### **8. Emerging Applications and Future Trends in ML-Based Decision-Making. The Role and Importance of Explainable AI (XAI) in Decision-Making**

As Machine Learning becomes increasingly integrated into organizations' decision-making processes, new directions in research and development are emerging that could radically transform how decisions are generated, explained, and executed.

One notable trend is the development of Explainable AI (XAI) technologies aimed at making models more transparent and easier to understand. This aspect is crucial in sectors such as healthcare, justice, or finance, where automated decisions must be justifiable to clients, authorities, or courts [19].

One of the most pressing current topics in the field of algorithm-assisted decisions is the need for transparency and understanding of how a Machine Learning model arrives at a certain conclusion. This emerging field is known as Explainable Artificial Intelligence (XAI) and seeks to answer the fundamental question: “Why did the model make this decision?”

Modern models, such as neural networks or ensembles, are often considered “black boxes.” Although high-performing, they cannot explain their decisions in human-understandable terms. This limits their trust and acceptability in critical contexts such as recruitment, healthcare, or judicial systems [20].

To address this issue, interpretability techniques are used, such as:

- **SHAP (SHapley Additive exPlanations)**, which evaluates the contribution of each feature to the prediction;
- **LIME (Local Interpretable Model-agnostic Explanations)**, which provides local explanations through simple models;
- **Surrogate Models**, interpretable models that mimic the behavior of a complex model.

These methods are integrated into modern platforms to provide auditability, traceability, and trust. At the same time, XAI is becoming increasingly important in light of regulations such as the GDPR and the AI Act, which require that automated decisions affecting individuals must be justifiable [21].

Thus, XAI is not just a technological enhancement, but an essential condition for ethical, responsible, and acceptable

artificial intelligence in organizational environments.

Organizations are also turning their attention to autonomous decision-making systems, capable not only of proposing actions but also of implementing them in real time. These systems are powered by Big Data infrastructures and continuous learning algorithms, with feedback from the operational environment. In logistics, for example, such systems can automatically reroute fleets based on real-time weather or traffic conditions.

Another emerging field is the use of generative models and conversational artificial intelligence (e.g., LLMs like ChatGPT) in decision support. These models not only analyze data but can also synthesize reports, explain results, or generate alternative scenarios, offering active support in strategic decision-making.

Moreover, there is growing interest in integrating ML into ESG (Environmental, Social, and Governance) initiatives to optimize energy consumption, detect unethical practices, or assess the social impact of decisions. In this way, algorithms become tools for responsibility, not just for efficiency.

In the near future, the emergence of hybrid decision-making ecosystems is anticipated, where artificial intelligence and the human factor collaborate continuously. Decisions will no longer be either automated or human-made but will result from a synergy in which algorithms learn from human judgment, and humans rely on the analytical power of ML.

## 9. Methodologies for Evaluating the Performance of Machine Learning Models

Evaluating the performance of machine learning models is crucial in the context of strategic decision-making, as it builds trust in the predictions and recommendations generated automatically. There are several standard methods and metrics used for

model evaluation, among which the most important are:

- Accuracy
- Precision
- Recall (Sensitivity)
- F1-Score
- Confusion Matrix
- ROC Curve and AUC Indicator (ROC-AUC)

In the following, we will detail these methods and their practical importance within decision-making processes.

#### *Accuracy*

Accuracy represents the percentage of total correct predictions made by the model out of the total number of predictions. It is the simplest and most intuitive metric, often used as a starting point in evaluating a model's performance [22]. However, accuracy can be misleading in the case of imbalanced datasets, where one class strongly outweighs the other [23].

#### *Precision*

Precision indicates the proportion of positive predictions that are actually correct, and it is essential in contexts where false positives are costly. For example, in credit risk assessment, high precision is vital to avoid unjustified loan approvals for high-risk individuals [24].

#### *Recall (Sensitivity)*

Recall shows the proportion of actual positive cases correctly identified by the model. It is critical in situations where false negatives are extremely important, such as in medical diagnosis or fraud detection, where identifying as many positive cases as possible is crucial.

#### *F1-Score*

The F1-Score is the harmonic mean between precision and recall. It is used to obtain a balanced evaluation, especially when there is a significant difference between the number of false positives and false negatives. Thus, the F1-Score is

suitable for situations where a balanced trade-off between precision and recall is desired.

#### *Confusion Matrix*

The confusion matrix is a table that provides a detailed view of the model's performance. It shows the number of correct and incorrect predictions split across classes (positive/negative), and is useful for understanding the specific types of errors the model makes. This method is widely used for in-depth performance diagnostics of models.

#### *ROC Curve and AUC Indicator (ROC-AUC)*

Receiver Operating Characteristic and the Area Under Curve indicator measure a model's ability to differentiate between positive and negative classes at various decision thresholds. An AUC value close to 1 indicates excellent performance, while a value of 0.5 indicates a model that is no better than random guessing. This method is widely used for the overall evaluation of predictive model performance across various industries.

### **10. The Impact of Big Data and Cloud Computing on Strategic Decision-Making Powered by Machine Learning**

In the digital transformation era, Big Data and Cloud Computing have become foundational pillars enabling organizations to implement machine learning (ML) at scale. These technologies significantly enhance the speed, accuracy, and adaptability of strategic decision-making processes. In this section, we explore their combined role in supporting data-driven decisions.

#### *The 5Vs of Big Data and Their Influence on Machine Learning*

Originally conceptualized through three dimensions (Volume, Velocity, and Variety), the Big Data paradigm has evolved to include five core



characteristics, known as the 5Vs [25], [26]:

- **Volume:** Refers to the massive scale of data generated daily from sensors, transactions, user interactions, and digital platforms. ML algorithms require large datasets to identify complex patterns and ensure high model generalizability.
- **Velocity:** Represents the speed at which data is generated, transmitted, and must be processed. Real-time data streams are critical in sectors such as finance or logistics, where immediate decisions are necessary.
- **Variety:** Describes the diversity of data formats-structured (e.g., databases), semi-structured (e.g., XML, JSON), and unstructured (e.g., video, audio, social media). Modern ML systems can integrate multimodal data sources to produce more robust insights.
- **Veracity:** Denotes the trustworthiness and quality of data. Inaccurate or biased data can lead to unreliable model predictions. Thus, data preprocessing and validation are essential stages in the ML pipeline.
- **Value:** Emphasizes the importance of extracting actionable insights from data. Data in itself holds no strategic utility unless transformed into knowledge that supports business goals [27].

Machine learning leverages these characteristics to derive predictive and prescriptive analytics, enabling organizations to anticipate trends, optimize resources, and identify risks proactively [28].

#### *The Role of Cloud Computing in Strategic Decision-Making*

Cloud computing provides a scalable, cost-effective, and accessible infrastructure for building and deploying machine learning solutions. Public cloud platforms such as Amazon Web Services (AWS), Google

Cloud Platform (GCP), and Microsoft Azure offer pre-built ML services and computing power that are essential for handling large-scale data processing tasks [29].

The main strategic benefits of cloud computing include:

- **Elastic scalability:** Organizations can dynamically allocate computing resources based on demand, reducing costs and improving performance.
- **Pay-as-you-go model:** Eliminates the need for capital-intensive hardware investments, allowing firms to experiment and scale ML projects efficiently.
- **Global accessibility:** Cloud infrastructure enables seamless collaboration and distributed access to models and data, supporting agile decision-making.
- **Tool integration:** Major cloud providers offer end-to-end platforms for ML workflows, data ingestion, model training, deployment, and monitoring, all in one environment [30].

Cloud computing lowers the entry barriers for implementing ML, making it accessible even to small and medium enterprises seeking data-driven competitiveness.

#### *The Synergy of Big Data and Cloud Computing in ML-Powered Decision-Making*

The convergence of Big Data and cloud infrastructure has created a powerful technological ecosystem that supports data-intensive machine learning applications. High-velocity and high-volume datasets can be stored and processed in real time using cloud-native tools, enabling faster insights and decision cycles.

A notable example is JPMorgan Chase, which utilizes cloud-based ML models to detect credit risk, prevent fraud, and optimize customer services based on predictive analytics. By integrating financial and behavioral data at scale, the

institution can adjust strategies dynamically in response to market signals [31].

In conclusion, the fusion of Big Data and cloud computing significantly amplifies the strategic value of machine learning. These technologies empower organizations to become more responsive, personalized, and data-driven in their operations, characteristics essential in an era of complexity and digital acceleration.

### 11. Legal and Regulatory Frameworks for the Use of Artificial Intelligence in Strategic Decision-Making

As artificial intelligence (AI) and machine learning (ML) systems become increasingly embedded in strategic decision-making processes, the need for clear legal and ethical governance is more urgent than ever. The absence of well-defined regulations can result in bias, discrimination, lack of accountability, and erosion of public trust. Thus, legal compliance is not just a technical concern but a strategic imperative.

#### *European Legislation: The AI Act and GDPR*

The European Union has taken a leading role in shaping legal frameworks for trustworthy AI. In April 2021, the European Commission proposed the Artificial Intelligence Act (AI Act), the first regulatory framework in the world to comprehensively govern AI technologies [32]. The Act introduces a risk-based classification system that divides AI systems into four categories:

- **Unacceptable risk** – applications that are strictly prohibited (e.g., subliminal manipulation, social scoring);
- **High risk** – systems used in sensitive domains such as finance, recruitment, and justice, which must comply with rigorous obligations including transparency, risk assessment, documentation, and human oversight;
- **Limited risk** – applications that require only basic transparency (e.g.,

chatbots that must disclose they are not human);

- **Minimal risk** – systems with negligible legal concerns, such as spam filters or video game AI.

High-risk systems, often used in strategic business decision-making, are subject to continuous monitoring, bias testing, and explainability requirements. Organizations must also implement mechanisms for human review and document the development and deployment lifecycle of ML models.

Complementing the AI Act, the General Data Protection Regulation (GDPR) already provides key protections regarding automated decision-making. Article 22 of GDPR gives individuals the right not to be subject to decisions based solely on automated processing that significantly affects them. It also grants them the right to receive a meaningful explanation of the logic involved [33]. These provisions compel organizations to use interpretable ML models or adopt Explainable AI (XAI) techniques to maintain transparency and accountability [34].

#### *Global Regulations and Emerging Ethical Standards*

Beyond the European context, other countries and organizations have also begun shaping AI governance. In the United States, while no federal AI law currently exists, the White House Blueprint for an AI Bill of Rights (2022) outlines five guiding principles for ethical AI [35]:

1. Safe and effective systems;
2. Protection from algorithmic discrimination;
3. Data privacy and agency;
4. Notice and explanation of AI decisions;
5. Human alternatives and fallback mechanisms.

Meanwhile, in the United Kingdom, the Information Commissioner's Office (ICO) has published detailed guidance on

algorithmic transparency, particularly for high-stakes applications like hiring, loan approval, or access to public services [36]. Global bodies such as the IEEE and the OECD have also issued AI ethics guidelines. Their key principles include fairness, explainability, human agency, technical robustness, and privacy protection [37], [38]. These frameworks aim to build public trust and encourage responsible AI deployment across sectors.

## **12. Strategic Implications of AI Regulation**

Beyond legal compliance, these regulatory frameworks influence how companies design and deploy AI models for strategic decision-making. Key organizational responsibilities include:

- Ensuring algorithmic transparency and interpretability, especially in regulated industries;
- Integrating human-in-the-loop controls for all high-risk decision processes;
- Maintaining comprehensive documentation of model training, validation, and deployment;
- Establishing AI governance structures, such as ethics boards and independent audits.

Failure to comply may lead to severe penalties. For example, under GDPR, violations can result in fines of up to €20 million or 4% of the global annual revenue, highlighting that AI compliance is a business risk as much as a legal one.

## **13. Conclusions**

In an organizational environment marked by uncertainty, volatility, and pressure for rapid decision-making, technologies based on Data Science and Machine Learning offer increasingly compelling solutions for the automation, optimization, and grounding of strategic decisions.

This paper highlighted how these technologies are transforming the decision-making process: from traditional descriptive reporting (Business Intelligence) to predictive and prescriptive

models capable of generating real-time insights. The analyzed case studies demonstrate the practical applicability of ML in fields such as finance, healthcare, public security, and digital commerce, offering not only operational efficiency but also long-term strategic value.

At the same time, the analysis of risks and limitations emphasized the need for a balance between performance and ethics. Issues such as lack of transparency, data bias, privacy, or lack of decision-making accountability can compromise not only the quality of the decision but also trust in the organization as a whole. In this context, initiatives like Explainable AI and the development of an algorithmic governance framework become essential for the sustainable use of these technologies.

A key direction for the future is not just improving the models themselves, but enabling organizations to build augmented decision-making processes, in which the human factor, contextual experience, and professional intuition are complemented—not replaced—by automated analysis. This requires investment not only in technology but also in organizational culture, education, and the development of advanced digital skills.

Furthermore, a deeper integration of ML into strategic planning systems is expected, through the generation of simulations, scenarios, and real-time risk assessments. Organizations will need to adopt an iterative decision-making model, in which outcomes are constantly tested and adjusted based on feedback and context. Decision-making will no longer be a fixed point, but an adaptive and intelligent process.

Ultimately, the use of ML in strategic decision-making should not be viewed as a technological revolution in itself, but as an evolution in the way organizations learn, adapt, and envision the future. Through responsible, transparent application focused on delivering real added value, these technologies can become a key pillar

of competitiveness and sustainability in the digital economy.

Beyond the obvious benefits brought by machine learning models, such as improved accuracy and decision efficiency, it is critical to understand how these models must be evaluated, regulated, and technologically supported to ensure long-term sustainability.

We observed that rigorous evaluation of algorithm performance is not only a technical necessity but also a strategic one. Metrics like precision, F1-score, and AUC directly contribute to the credibility of automated decisions and the trust of stakeholders in these systems. Additionally, Big Data and Cloud Computing ecosystems provide the essential infrastructure for scalability, accessibility, and responsiveness in implementing ML models across organizations.

At the same time, international regulations, especially the AI Act and GDPR, define the ethical and legal boundaries of automated decision-making. As a result, companies can no longer overlook requirements related to transparency, explainability, and data protection. Compliance with such regulations not only mitigates legal risks but also becomes a competitive differentiator in an environment where technological accountability is increasingly demanded.

In conclusion, the use of machine learning in strategic decision-making should not be viewed solely through the lens of algorithmic performance, but rather as a process governed by ethical principles, legal frameworks, and robust infrastructure. The success of these technologies ultimately depends on achieving a balance between innovation and responsibility.

## References

- [1] Provost, F., & Fawcett, T. (2013). *Data Science for Business*. O'Reilly Media.
- [2] Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media.
- [3] Shmueli, G., Bruce, P. C., Gedeck, P., & Patel, N. R. (2020). *Data Mining for Business Analytics: Concepts, Techniques, and Applications in Python*. Wiley.
- [4] Kelleher, J. D., & Tierney, B. (2018). *Data Science*. MIT Press.
- [5] Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media.
- [6] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning*. Springer.
- [7] Wixom, B., & Watson, H. (2010). *The BI-Based Organization*. *International Journal of Business Intelligence Research*, 1(1), pp.13–28.
- [8] Waller, M. A., & Fawcett, S. E. (2013). *Data Science and Predictive Analytics in Supply Chains*. *Journal of Business Logistics*, 34(2), pp.77–94.
- [9] Chen, H., Chiang, R. H. L., & Storey, V. C. (2012). *Business Intelligence and Analytics: From Big Data to Big Impact*. *MIS Quarterly*, 36(4), pp.1165–1188.
- [10] Shmueli, G., et al. (2020). *Data Mining for Business Analytics*. Wiley.
- [11] Venkatapathi, S., & Vishnuvardhan, S. (2025). *Leveraging Azure AI and Machine Learning for Predictive Analytics and Decision Support Systems*. [Online Manuscript/Whitepaper].
- [12] O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown Publishing Group.
- [13] PwC (2022). *AI Predictions Report*.  
IEEE (2023). *Technology Survey on AI Trust*.  
McKinsey & Company (2021). *Global AI Adoption Study*.
- [14] JPMorgan Chase & Co. (2022). *AI and Machine Learning in Financial Services*.

- [15] Ross, C., & Swetlitz, I. (2017). *IBM pitched its Watson supercomputer as a revolution in cancer care. It's nowhere close. STAT News.*
- [16] Amazon (2021). *How Amazon personalizes the shopping experience. AboutAmazon.*  
<https://www.aboutamazon.com>
- [17] Metropolitan Police (2020). *Live Facial Recognition – Legal Framework and Results.*
- [18] Gomez-Uribe, C.A., & Hunt, N. (2015). *The Netflix Recommender System: Algorithms, Business Value, and Innovation.*
- [19] Mozur, P. (2018). *Inside China's Dystopian Dreams: AI, Shame and Lots of Cameras. The New York Times.*  
<https://www.nytimes.com/2018/07/08/business/china-surveillance-technology.html>
- [20] Rashid, M.H.O., Hoque, M.R., & Sarkar, M. (2025). *Explainable AI In E-Commerce: Enhancing Trust And Transparency In AI-Driven Decisions. Innovatech Technology & Engineering Journal.*
- [21] Kaur, K., Agrawal, A., & Kaur, H. (2024). *Explainable AI in Biometrics: A Novel Framework for Facial Recognition Interpretation. IEEE Conference Proceedings.*  
<https://ieeexplore.ieee.org/document/10881703>
- [22] Aluvalu, R., Mehta, M., & Siarry, P. (2024). *Explainable AI in Health Informatics.* Springer.  
<https://link.springer.com/content/pdf/10.1007/978-981-97-3705-5.pdf>
- [23] Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow.* O'Reilly Media.
- [24] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning.* Springer.
- [25] Provost, F., & Fawcett, T. (2013). *Data Science for Business.* O'Reilly Media.
- [26] IBM. (2015). The Four Vs of Big Data. IBM Big Data & Analytics Hub. Retrieved from:  
<https://www.ibm.com/think/topics/big-data>
- [27] Laney, D. (2001). 3D Data Management: Controlling Data Volume, Velocity, and Variety. Meta Group (later acquired by Gartner).
- [28] Marr, B. (2016). Big Data in Practice: How 45 Successful Companies Used Big Data Analytics to Deliver Extraordinary Results. Wiley.
- [29] Waller, M. A., & Fawcett, S. E. (2013). Data Science and Predictive Analytics in Supply Chains. *Journal of Business Logistics*, 34(2), 77–94.
- [30] Venkatapathi, S., & Vishnuvardhan, S. (2025). Leveraging Azure AI and Machine Learning for Predictive Analytics and Decision Support Systems. Online Manuscript.
- [31] Amazon Web Services (AWS). (2024). AWS for Machine Learning. Retrieved from:  
<https://aws.amazon.com/machine-learning/>
- [32] JPMorgan Chase & Co. (2022). AI and Machine Learning in Financial Services. Corporate Publication.
- [33] European Commission. (2021). Proposal for a Regulation on a European Approach for Artificial Intelligence (AI Act). Retrieved from:  
<https://eur-lex.europa.eu>
- [34] GDPR – Regulation (EU) 2016/679 of the European Parliament. Retrieved from: <https://gdpr.eu>
- [35] Aluvalu, R., Mehta, M., & Siarry, P. (2024). *Explainable AI in Health Informatics.* Springer.  
<https://link.springer.com/content/pdf/10.1007/978-981-97-3705-5.pdf>
- [36] White House. (2022). Blueprint for an AI Bill of Rights. Retrieved from:  
<https://bidenwhitehouse.archives.gov/ostp/ai-bill-of-rights/>
- [37] ICO. (2022). Explaining Decisions Made with AI. UK Information

Commissioner's Office. Retrieved from: <https://ico.org.uk>

[38] IEEE. (2023). Technology Survey on AI Trust. Retrieved from: <https://ethicsinaction.ieee.org>



**Elena-Luiza JALEA** completed her undergraduate studies at the Faculty of Economic Cybernetics in 2023 and graduated from the master's program in Databases and Business Support Systems at the Bucharest University of Economic Studies in 2025. She has completed an internship as a Risk Data Scientist, where she worked on data modeling and analysis projects. Her main areas of interest include data science, machine learning, and web application development, with a particular focus on building intelligent systems to support decision-making.



**Robert-Marian VLĂDUȚĂ** completed his undergraduate studies at the Faculty of Economic Cybernetics in 2023 and graduated from the master's program in Databases and Business Support Systems at the Bucharest University of Economic Studies in 2025. He currently works as a Lead Data Scientist, coordinating data-driven projects focused on risk modeling and advanced analytics. His main areas of interest include machine learning, artificial intelligence, data science, data analysis, and game development.

## Comparative Aspects of E-learning Platforms in the Field of Foreign Languages

Georgiana VOICA

Bucharest University of Economic Studies

Faculty of Cybernetics, Statistics and Economic Informatics

Bucharest, Romania

[voicageorgiana20@stud.ase.ro](mailto:voicageorgiana20@stud.ase.ro)

*Novelty has become something natural, and innovation has become a norm, which fuels the fierce desire to constantly identify new alternatives. This paper explores the characteristics, advantages and limitations of e-learning platforms dedicated to foreign language acquisition. A comparative analysis was made between platforms such as Duolingo, Babbel, Busuu, Mondly, Seedlang, Drops and DW Learn German, which is based on various criteria such as gamification, vocabulary, grammar or AI. The research also includes an empirical component, based on an user questionnaire that aims to discover the way that the MALL (mobile-assisted language learning) apps are perceived by the public. Furthermore, the study discusses the challenges of these platforms and outlines future directions such as AI-driven personalization, immersive technologies (AR/VR) and integration into classic education systems.*

**Keywords:** e-learning, languages, MALL, AI, gamification

### 1 Introduction

Digitalization has also "infiltrated" the field of education, causing a strong change in the way the learning process is perceived. Classical methods have become obsolete in a very short time, and digital platforms have become a preferred alternative for the majority, replacing key elements of conventional didactics such as books and teachers.

The trend of using this new variant of study was also manifested in the case of foreign languages, because people perceived the emergence of these IT solutions as the beginning of a new era, one of convenience and flexibility. Thus, barriers related to interlingual and intercultural communication, as well as those related to distance, were removed, allowing any person to learn a new language, without having to travel or physically interact with other people. The focus of this study falls on the German language, being one of the most spoken languages in Europe and which has major economic importance at a global level.

The paper aims to analyze the benefits of e-learning applications for learning German, which benefit from a strong diversity among dedicated digital resources. The study focuses on representative platforms such as Duolingo, Mondly, Babbel, Busuu, Seedlang, Drops and DW Learn German, each offering different approaches. The entire research will be divided into several chapters, starting with the presentation of the theoretical framework, followed by the presentation and comparison of the platforms based on several criteria, limitations, integration in formal education, and finally conclusions will be presented regarding trends in this field.

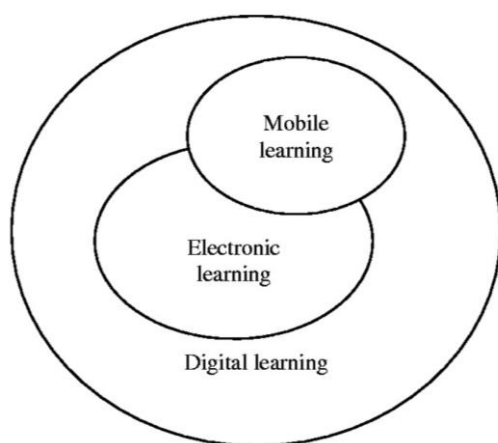
### 2. Theoretical context

**E-learning** is a term that has gained weight in the academic context, taking over learning centers with aplomb and gradually integrating into the teaching process, becoming essential for the transmission of information. According to the definition, e-learning means electronic learning and involves the use of electronic



resources such as the Internet, communication networks and interactive software to access the materials made available, to accumulate knowledge, but also to develop through the learning process. [1]

The term hides under its "umbrella" a multitude of components, which thus offer a variety of ways in which e-learning can be used. Here we can talk about asynchronous learning (users use the materials at their own pace) or synchronous learning (that relating to real-time interaction). We can also extrapolate the term e-learning, which has currently acquired new valences through the increasingly common use of the notions of m-learning (learning through mobile applications) and d-learning (the entire digital learning process). It is known that d-learning has gradually begun to replace the general term e-learning, incorporating all the tools and processes necessary for online study, moving from moodle platforms to advanced mobile applications. MALL (mobile assisted language learning) applications are the most common and those that offer the best results in terms of the degree of efficiency in learning a foreign language. [2]



**Fig. 1.** The relationship between m-learning, e-learning, d-learning

## 2.1 Advantages

The main advantages provided by a MALL application are highlighted in various studies, which presents how online learning software has eliminated the space

limitations imposed by conventional learning methods. The variety of areas from which students can access and deepen foreign language lessons has overcome the restrictions of the past, as they are no longer obliged to support them only from certain classrooms or rooms equipped with appropriate equipment. The working area has been massively reduced, with the emphasis on mobility and flexibility, which has generated an increase in motivation. Thus, anyone eager to master a new language can do so simply by using an application installed on their mobile phone. Regardless of where they are, learners can easily access the resources provided, without feeling pressed for time or having to travel to a specific location [3].

The idea of mobility basically refers to the independence of time and space of learning through mobile phones, referring to them as "24/7, barrier-free education centers". This produces benefits especially among students who do not perform so well in the classroom, but manage to find their motivation when using m-learning from home or even in informal activities, because they study at their own pace, without being compared to others. The results show that MALL technologies produce a significant increase in the degree of autonomy, but also in self-confidence and give learners a sense of achievement through their own strengths [4].

People who choose to use mobile applications to learn a new language are primarily attracted to gamification. This term refers to the integration of games into educational activities. Tasks initially perceived as difficult are transformed with the help of these didactic devices into enjoyable activities that stimulate cognitive capacity and reduce stress. Concrete examples that justify the attraction that learners show towards this type of applications are word games that contribute to the development of vocabulary or simulation games that improve pronunciation [5].

We can see how all these improvements in functionality have reduced the impediments to language acquisition through speech, as you can now interact with virtual chatbots, which simulate human conversations by integrating AI tools, or even with real native speakers. Contact with people from different cultures strengthens interlingual dialogue skills, as understanding customs and traditions leads to better communication, one based on respect, harmony and unity.

### 3. Presentation of the platforms analyzed

As I said in the previous article, the most famous language learning application globally is Duolingo, offering courses in over 30 languages, and can be used both in the web and mobile versions. Easy to use, interactive and extremely attractive, this mobile solution is chosen by people of all ages because it manages to attract them through simplicity, but also through parasocial marketing. Although in this field, capability is determined by how practical the lessons are and how well chosen the exercises are (these are divided into different themes), there is a growing trend in the social impact that these applications have [6].

Duolingo has the same technical functions as its competitors, but it stands out for managing to "sell" itself through social media promotions (funny posts that give the brand an almost human valence, with which most users identify; informal interactions of the page with other users of social media applications)[7]. In addition, the rewards offered at the end of the lessons are represented by the idea of *gamification* (badges, points in a ranking, tracking statistics), which arouses a magnetic interest among learners [8].

**Babbel** differs from Duolingo primarily in its scientific profile, the application being designed by linguistic experts. This additionally includes the option to participate in meditations supported by accredited people, the scheduling being

done within the application, depending on criteria such as availability, number of places, language level [9].

**Mondly** is a paid application that structures the lessons into various essential categories in mastering a foreign language (necessary for carrying out everyday conversations), and each chapter consists of several lessons. We also find vocabulary pages and a dialogue page (where you can converse with a chatbot to practice). Like Babbel, there is also information related to statistics and progress within the application.

**Seedlang** is an application with a completely different approach from those on the market. It presents lessons in the form of stories, told by native speakers, which capture different contexts that can occur in everyday life. The emphasis is on speaking and listening, so users can hear their own pronunciation and compare it with that of another German speaker in order to improve it. Also, another interesting functionality is represented by flashcards, which users can build themselves to practice different terms. Grammar and vocabulary are also present, with a dedicated button, which by pressing which, learners can discover information related to pronunciation, concrete examples of use and helpful explanations. Another interesting component is represented by the trivia section, where there are mini-games dedicated to learning new words and which aim to motivate through competition with other users.

**Busuu** is similar to Seedlang, but has a few things that make it stand out. While the lesson structure is still based on videos with native speakers, there is also an integrated AI component that analyzes the pronunciation part and provides feedback. The community page allows members to post recordings of various phrases to improve their pronunciation, through comments left by other users.

Another popular app is **Drops**. It stands out for its gamification-based approach, with users being encouraged to go through as

many topics as possible through games. They must match the word (noun, adjective, verb) to the corresponding image that describes it. The solution is built to be in line with the fast-paced society we live in, with the expectation that the user will progress at a leisurely pace, according to their own schedule. To draw attention to itself, the app sends reminders and motivates its learners through daily streaks. The last app reviewed was **DW Learn German**. This app is a reference point for all German speakers, as it is built around the idea of Deutsche Welle, which started as a radio station, which gained notoriety, and over time expanded to a German-language news website, which became recognized worldwide for its verticality and has now also come to offer German language courses (both in web and mobile form). Each chapter aims to present the lives of different characters, and can be watched as a mini-series, which incorporates everyday language, making it particularly easy for users to feel like they are part of the story. At the end of each episode, there are lessons to fix vocabulary and grammar, and users can even hear their own pronunciation.

#### 4. Comparative analysis

E-learning platforms for learning German differ significantly from each other, both in terms of structure and the technologies used. This analysis was based on the following **criteria**: type of content, feedback, grammar, vocabulary, level of human interaction, gamification, accessibility and integration of artificial intelligence.

##### 4.1 Theoretical foundation of comparison criteria

In order for the analysis of these language learning solutions to be consistent with the requirements of scientific research, we chose as basic “measures” relevant pedagogical and technological criteria, which reflect the learning process of a standard user from several angles. Thus, the comparison made is rigorous,

especially because the attention is polarized on different aspects that can be addressed when we talk about an examination of an information technology in the educational field.

The criteria selected in this paper – lesson content, feedback, vocabulary, grammar, human interaction, gamification elements, price and use of artificial intelligence – are frequently used in specialized studies on e-learning, and for this reason I have chosen to briefly argue their use, as well as a brief definition [10] [11].

##### a) Lesson content

The content of a platform represents a fundamental step in the learning process. The way it is structured plays a decisive role in guiding a learner's progress, which makes the lesson preparation process the basic stage. According to some studies, teaching materials must be relevant, motivating and adapted to the real communication needs of learners [12].

##### b) Feedback

Feedback is another extremely important component in the process of knowledge accumulation. In online environments, it can be immediate, automated, or personalized. Thus, formative feedback contributes to improving performance by identifying gaps and offering concrete suggestions. Platforms that provide detailed feedback (not just binary, right/wrong) favor self-regulated learning.

##### c) Vocabulary

When we want to learn a new foreign language, the first step we need to take is to assimilate new words. Vocabulary thus becomes a pillar that supports a language, and the newer terms we learn and the more often we practice, the closer we are to achieving our goal.

##### d) Grammar

Vocabulary would not be sufficient without the grammar component, which can be presented: explicitly (through rules) or implicitly (through examples). That being said, a combination of deductive and inductive learning is most effective, especially in technology-assisted

environments. Platforms that offer clear explanations and contextualized application exercises are more effective [13].

#### e) Human interaction

We often hear that a language must be practiced in order not to be forgotten. And a very important aspect in this regard is the dialogue with native speakers. They help us to learn the correct pronunciation, vocabulary, and grammar more quickly. Applications that manage to introduce this component are becoming increasingly sought after today.

#### f) Gamification

Gamification refers to the integration of “game” elements (badges, levels, points) into the educational process. It is shown that these mechanisms can increase user motivation and engagement. In language learning, gamification helps create a

rhythm and routine for learning, through daily challenges and visual feedback of progress [14].

#### g) Price and accessibility

Another very important criterion (and one that attracts users' attention) is the free versions of the applications (the freemium model) and how much they can offer compared to paid subscriptions. Democratized access to education is a key principle in e-learning [15].

#### h) Artificial intelligence

AI has become the most sought-after functionality of contemporary technology. By personalizing the learning process (recommendations), voice recognition or providing dynamic feedback, AI has managed to become an indispensable component, without which learning would not be so fun.

**Table 2.** Comparative analysis of the studied platforms

Platform	Content	Feedback	Grammar	Vocabulary	Human Interaction	Gamification	Price	AI Integrated
<b>Duolingo</b>	Interactive lessons	Automatic	Summary	Extended	Miss	Yes	Free & subscription	Yes (prediction /vocal feedback)
<b>Babbel</b>	Interactive lessons, dialogue	Customized	Detailed	Moderate	Yes (chat)	No	Subscription	Yes (feedback)
<b>Mondly</b>	Interactive lessons, AR, chatbot	Automatic	Summary	Extended	Chat	Yes	Free & subscription	Chatbot AI, (AR)
<b>Busuu</b>	Video lessons, conversations	Human, Automatic	Detailed	Extended	Natives	No	Free & subscription	Yes (recommendations)
<b>Seedlang</b>	Video lessons	Human (partially)	Developed	Really extended	Natives	No	Free & subscription	Missing
<b>Drops</b>	Games	Automat	Weak	Visual	Miss	Yes	Free & subscription	Missing
<b>Dw learn german</b>	Video lessons, exercises	No	Developed	Extended	Miss	No	Free	

## 4.2 Content

All the applications presented offer various language levels, starting from A1 and even up to C2, but Seedlang, Drops or Mondly place the predominant emphasis on the

beginner and intermediate level, insisting on notions related to vocabulary and expressions encountered in everyday conversations. The content is varied, most of which contain interactive lessons with

well-structured exercises. Seedlang is distinguished by videos featuring native speakers, while Drops uses a minimalist approach, namely games of associating certain terms with visual elements to make it easier to remember. Mondly is the only application that stands out by integrating augmented reality, but this option can only be used by people who have purchased VR (virtual reality) glasses.

#### 4.3 Feedback, grammar and vocabulary

Regarding the feedback part, **Babbel** and **Busuu** offer the most developed solutions, because they manage to add in addition to the automatic feedback the possibility of evaluation by people who already master the German language. **Duolingo**, although it is the most famous application, is quite limited in this aspect, offering only automated feedback, lacking a substrate.

Although grammar should be an essential component, **Drops** completely excludes this criterion and focuses exclusively on the visual memorization of terms. On the other hand, **Babbel** and **DW Learn German** manage to include all the explanations related to the lexical and morphological parts, necessary for the correct speaking of a language. The vocabulary part is covered by all platforms, but **Drops**, **Mondly** and **Seedlang** emphasize this aspect, being the most suitable for the rapid assimilation of words.

#### 4.4 Interaction and gamification

It can be noted how platforms like **Busuu** (through native correction) and **Babbel** (which offers live sessions or simulated conversations) manage to use human interaction. **Duolingo** and **Mondly** compensate for the lack of real human interaction through gamification elements: points, levels, daily progress and rankings. This approach helps maintain motivation, especially for beginners.

#### 4.5 Integration of artificial intelligence and machine learning

One of the defining aspects that differentiates these platforms is the degree

to which they integrate artificial intelligence (AI) and machine learning (ML) technologies. Platforms like **Duolingo**, **Babbel**, **Mondly**, and **Busuu** use AI to personalize the learning experience, adapting content and exercises based on the user's performance and habits.

For example, Duolingo uses ML algorithms to predict words that are likely to be forgotten, and **Babbel** offers a "revision manager" that adjusts content based on individual difficulty. **Mondly** even introduces an AI chatbot, along with augmented reality, to create an interactive conversational environment. **Busuu**, on the other hand, combines technology with a social component, providing intelligent recommendations and corrections from native speakers.

In contrast, platforms like **Seedlang**, **Drops**, and **Deutsche Welle** do not integrate AI significantly, relying on fixed content, without automatic adaptation or analysis of user performance. Thus, the difference between traditional and AI-based platforms becomes visible in the level of personalization and efficiency of learning.

#### 4.6 Integration of artificial intelligence and machine learning

Most platforms offer interfaces in Romanian or English, but some, such as **Seedlang**, are only available in English, which may limit access for certain categories of users. **Deutsche Welle** has an interface available in several languages, but it requires a greater degree of autonomy in learning.

In terms of financial accessibility, **Duolingo** and **Deutsche Welle** are the most accessible – the first through the freemium model (i.e. both free and with the possibility of purchasing a subscription), and the second completely free. **Babbel** and **Busuu** require monthly subscriptions for full access, which may influence the choice of users depending on their budget.

## 5. Empirical analysis

### 5.1. Empirical study on user perception of e-learning platforms

To strengthen this analysis, an online questionnaire was developed to assess users' perceptions of e-learning platforms used in language learning. The instrument

included closed-ended questions (scale from 1-5) and open-ended questions, and its distribution was carried out through social networks and university groups. The questionnaire was completed by 23 participants

**Table 3.** Structure of questionnaire

Question	Answer options
1. How old are you?	<18 18–25 26–35 36–45 > 45
2. What is the foreign language you mainly learn through platforms?	English German Spanish French Other
3. What language learning platforms have you used? (you can check more than one)	Duolingo Babbel Mondly Busuu Drops Seedlang DW Learn German Other
4. How frequently do you use e-learning platforms for foreign languages?	Daily Several times a week Once a week Rarely I don't use anymore
5. On a scale of 1 to 5 (1 = very poor, 5 = excellent), how would you rate your preferred platform on the following aspects?	a. Clarity and structure of lessons b. Quality of vocabulary exercises c. Quality of grammar exercises d. Possibility to receive useful feedback e. Human interactivity (chat, corrections from users) f. Gamification elements (badges, points, levels, etc.) g. Use of artificial intelligence (AI) for personalization h. Value for money i. Which platform do you find most effective for learning a foreign language? Why? j. Have you had any difficulties using any platform? If so, which one and what type of difficulty? k. What would you like these platforms to improve?

### 5.2. Results and their interpretation

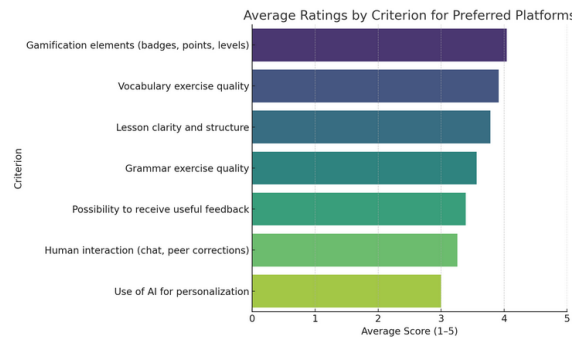
The majority of respondents (43.5%) fall into the 18–25 age group, followed by the 26–35 age group (30.4%). This confirms that the core audience of platforms like

*Duolingo*, *Babbel* or *Seedlang* is young and digitally active.

In terms of languages learned, English (34.8%) and German (26.1%) are dominant, followed by Spanish and French. This profile is consistent with the

languages predominantly offered by the platforms analyzed in the paper.

**Duolingo** is by far the most used platform – mentioned by 82.6% of respondents – followed closely by **Babbel**, **Busuu**, **Seedlang** and **DW Learn German**, each with around 8.7%. This can be attributed to the free availability, attractive interface and high notoriety of the application.



**Fig. 2.** Results of the questionnaire

In terms of frequency of use, only 21.7% use the platforms daily, while 39.1% state that they use them less frequently, which may reflect a lack of consistency or difficulties in maintaining motivation without a formal framework.

Participants rated their preferred platform on a scale from 1 to 5 according to several aspects. The average results obtained are as follows:

**Table 4.** Criteria and average score

Criteria	Avg. Score
Gamification (badges, points, levels)	<b>4.13</b>
Value for money	<b>4.00</b>
Vocabulary exercise quality	3.96
Lesson clarity and structure	3.87
Grammar exercise quality	3.65
Useful feedback	3.52
Human interaction (chat, peer corrections)	3.35
Use of AI for personalization	3.00

The platforms are perceived positively especially in terms of **gamification** and **value for money**, which confirms the importance of motivational design and financial accessibility. Playful elements

seem to be a determining factor in attracting and retaining users.

**Vocabulary** and **clarity of lessons** also receive high scores, indicating a good pedagogical organization of the content. In contrast, **human interactivity** and the **use of AI** received lower scores, which may indicate either a lack of these features or a superficial use of them by users. It is possible that the platforms do not offer enough opportunities for authentic conversation or advanced personalization, two important aspects for linguistic progress.

In addition to quantitative data, respondents were invited to provide personal opinions on the effectiveness of the platforms used, the difficulties encountered and possible directions for improvement.

### The platforms considered the most effective

Most respondents mentioned Duolingo as the most effective platform, citing reasons such as:

- user-friendly interface,
- clarity in the organization of lessons,
- variety of exercises,
- “fun” or “interactive” character.

One respondent states: “*Duolingo is an interactive, colorful, easy-to-access platform that keeps you motivated.*”

Other users appreciated Seedlang, especially for its video lessons and correct pronunciation: “*Seedlang seems to me the best because it personalizes well and emphasizes pronunciation.*”

### 5.3 Difficulties encountered

Most respondents stated that they did not have significant difficulties. However, a few recurring aspects are worth mentioning:

- Restrictions in the free version of platforms such as **Babbel** or **Mondly**: “*Babbel, because of the pricing system, sometimes it is frustrating that I cannot access all the content.*”



- Lack of real human interaction – which corroborates the lower score given to this criterion in the quantitative assessments.

#### 5.4. Suggestions for improvement

The responses were diverse, but can be grouped into a few key categories:

- More human interactivity / real conversations: *"I would like to be able to communicate live with other users who speak the same language."*
- Better adaptation to the level of difficulty: *"I would like it to have different levels of difficulty and personalized exercises."*
- Diversification of exercises and content: *"More variety in exercises, maybe on topics of current interest."*
- Improvement of simulated conversations: *"Conversations should be closer to real, natural ones."*

#### 6. Current limitations and future development directions of platforms

Although e-learning platforms have taken on the study of a foreign language with aplomb in recent years, both on the technological and educational fronts, we can still talk about limitations related to functionalities and content. It is necessary to recognize these obstacles and study them so that in the future we can offer more efficient solutions that are specifically oriented to each learning process.

##### 6.1. Suggestions for improvement

Following the analyzed questionnaire, we can notice some directions given by the learners themselves. A first feedback refers to the lack of authentic human interaction. As previously mentioned, a foreign language cannot be fully known if it is not spoken, and interaction with native speakers or even with other people who know the language plays a major role in cementing the verbal component. Although

there are also applications that manage to partially integrate this component (through videos, artificial intelligence, feedback), they often require a paid subscription, and the free version is seen as deficient and needs to be expanded in order not to jeopardize the user's progress, when we refer to creating a connection with a person who uses another language through conversation.

Another difficulty may also be imposed by the educational path. Although adaptation algorithms are used within certain platforms, either personalization is treated superficially or must be accessed with money. The content often ends up being standardized, sequential, focused on following a pre-established template and not taking into account the particularities of the learning style or even the individual objectives of the learner. This issue was raised by several respondents, who expressed their desire to have a more customized learning path.

As we have already mentioned, financial accessibility hinders the educational process, and essential functionalities can only be accessed by paying. It is understandable that platforms need many resources to create complex exercises that adapt to each person's learning style, but it is necessary to understand that when we talk about studying a foreign language, the practical (conversational) component is also needed. Survey respondents reported this aspect in the case of Babbel and Mondly, indicating a frustration related to the limitations of the free versions. Motivation also represents another major obstacle, since there is a tendency among users to abandon the progress made up to the present moment, if they do not feel external pressure, as happens in actual courses. This can come from teachers, deadlines or even from the colleagues involved. Although gamification largely contributes to keeping users active, it cannot fully replace interpersonal relationships.

## 6.2. Development directions

In order to overcome these barriers, it is necessary to establish concrete and innovative solutions from the point of view of the techno-educational process. Therefore, we have created this graphic that reflects the first steps that need to be taken in this direction.

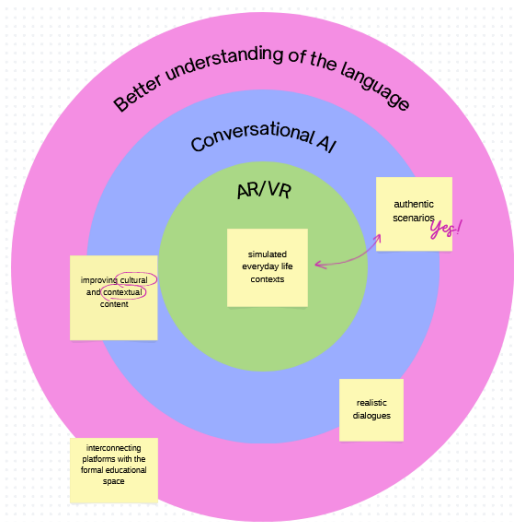


Fig. 3. Proposed solutions

To minimize the lack of interaction with other people, augmented reality (AR) and virtual reality (VR) can be implemented. This gives users the chance to practice their language in various simulated contexts such as a conversation at the train station, in the park, at the supermarket, etc. These technologies are already being researched as possible options to overcome the current barriers of language learning platforms [16].

Of course, this can only be possible through the use of artificial intelligence. Another branch of AI could be supported by the integration of a conversational AI, with the help of which learners can interact with a virtual assistant that simulates realistic dialogues. As many platforms as possible should consider integrating a chatbot to allow learners to receive dynamic feedback.

In addition, the solutions presented above lead to an important part that is often omitted, which consists of increasing the cultural and contextual content, since it is important for users to understand how

people actually speak depending on the situation, interlocutor and culture, not just to memorize some expressions or grammatical rules.

Finally, placing these functionalities in an academic context can contribute to much deeper learning, and used as adjacent tools, they become an interactive means for deepening foreign languages.

## 7. Integrating e-learning platforms into formal education

E-learning platforms should not be restricted to being used only in leisure time, but should also be introduced into formal education. Thus, these instruments who were initially used outside the institutional framework, could now complete and modernize the traditional ways of teaching foreign languages in schools and universities.

### 7.1. Arguments for integration

Comparing the platforms presented in this study, Duolingo, Babbel, Busuu or DW Learn German offer a wide range of functionalities that can extend the classic curriculum:

- Multiple accessibility (via phone, tablet, or PC)
- Interactive and engaging exercises that boost motivation
- Immediate feedback that supports personalized and individualized learning
- Adaptability to the learner's level, including automatic review sessions

In an educational setting, these foreign language learning platforms can be used as:

- Methods for reinforcing lessons taught in class
- Tools for ongoing formative assessment
- Resources for independent study

This blended learning model is also supported by numerous studies, which highlight the pedagogical benefits of combining classic teaching methods with

new digital solutions, specially designed to serve this purpose [10].

### 7.2. Examples of good practices

There are already initiatives where platforms have been successfully integrated into formal education:

- In some schools in Germany and Austria, **DW Learn German** is used to support migrant students in the process of linguistic integration [17].
- **European Erasmus+ projects** have used mobile applications for language learning in intercultural contexts, emphasizing autonomy and gamification [18].
- Some universities offer credits for completing digital modules in Duolingo or Memrise, as support for foreign language courses [19] [20].

### 7.3. Challenges and limits of integration

Although the advantages are numerous, integrating these platforms into everyday teaching and learning system also comes with challenges:

- Uneven quality of content: not all platforms follow the same curricular structure
- Lack of digital training for teachers: not all teachers are prepared to use these tools effectively in a pedagogical way.
- Differences between platforms: some focus on vocabulary, others on grammar, and others on conversation, which makes standardization difficult.

### 7.4. Insights and recommendations

For e-learning platforms to become an active part of classical education, it is necessary to:

- Careful selection of applications appropriate to the level and objectives of the course;
- Training teachers in their pedagogical use;
- Adapting the platform content to the requirements of the school curriculum;

- Integrating them into a well-planned hybrid system (blended learning).

Thus, instead of competing with traditional methods, digital platforms can become an additional educational partner, contributing to more interactive, motivating learning adapted to the current needs of students.

## 8. Conclusions

The comparative study highlights the diversity and complexity of e-learning platforms dedicated to learning German. Each application analyzed presents specific strengths, addressing different needs and learning styles. Platforms such as **Babbel** and **Busuu** stand out for their structured content, personalized feedback and extensive coverage of language levels, while **Duolingo** and **Mondly** offer a gamified, accessible and engaging experience, based on modern technologies such as AI and ML. Although platforms such as **DW Learn German**, **Drops** or **Seedlang** do not actively integrate intelligent technologies, they offer authentic content and valuable resources for strengthening language skills.

This paper not only summarized the *advantages* of the listed platforms, but also the current *limitations* that hinder user progress, such as the lack of human interaction, fluctuating motivation, and the barriers raised by paid subscriptions.

The applied *questionnaire* was a good way to confirm these hypotheses, and user suggestions should be listened to, as more rigorous AI integration is the clear path that language learning applications should follow.

Therefore, choosing an *effective platform* depends on the user's profile, goals and learning style preferences. At the same time, the integration of artificial intelligence remains an essential criterion in differentiating modern platforms, with significant potential to personalize and streamline the language learning process.

## References

- [1] M. Ally, "Foundations of educational theory for online learning," *The theory and practice of online learning*, pp. 15-44, 2008.
- [2] K. B. Sujit, W. Marguerite and B. Paul, "E-learning, M-learning and D-learning: Conceptual definition and comparative analysis," *E-Learning and Digital Media*, vol. 15, no. 4, pp. 191-216, 2018.
- [3] Elena, D. Svetlana and M. K. Ibrahim, "Using M-Learning Technology in Teaching Foreign Languages: A Panacea during COVID-19 Pandemic Era," *International Journal of Interactive Mobile Technologies*, vol. 15, no. 15, 2021.
- [4] J. Kacetl and B. Klimova, "Use of Smartphone Applications in English Language Learning—A Challenge for Foreign Language Education," *Education Sciences*, vol. 9, no. 3, p. 179, 2019.
- [5] R. Gafni, D. B. Achituv and G. Rahmani, "Learning Foreign Languages Using Mobile Applications," *Journal of Information Technology Education*, vol. 16, no. 1, pp. 301-317, 2017.
- [6] G. Voica, "M-Learning Solution For Learning A Foreign Language," 2023.
- [7] N. Laanti, "Parasocial relationships between businesses and customers as a way of marketing," *International Business*, p. 46, 2023.
- [8] K. Teske, "Duolingo," *CALICO Journal*, vol. 34, no. 3, pp. 393-401, 2017.
- [9] S. Beheshti, "Babbel: A Mobile Language Learning App," *TESL Reporter*, vol. 51, no. 1, pp. 109-121, 2018.
- [10] N. Hockly, "Blended Learning," *ELT Journal*, vol. 72, no. 1, pp. 97-101, 2018.
- [11] Kukulska-Hulme, "Will mobile learning change language learning?," *ReCALL*, vol. 21, no. 2, pp. 157-165, 2009.
- [12] J. C. Richards, *Curriculum Development in Language Teaching*, Cambridge University Press, 2001.
- [13] Rod, "Current Issues in the Teaching of Grammar: An SLA Perspective," *TESOL Quarterly*, vol. 40, no. 1, pp. 83-107, 2006.
- [14] S. Deterding, D. Dixon, R. Khaled and L. Nacke, "From Game Design Elements to Gamefulness: Defining Gamification," in *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments*, 2011.
- [15] N. Selwyn, *Education and Technology*, 2016.
- [16] R. Godwin-Jones, "Using mobile technology to develop language skills and cultural understanding," *Language Learning & Technology*, vol. 22, no. 3, pp. 104-120, 2018.
- [17] Welle, "Learning and Teaching German," <https://learngerman.dw.com/en/learning-and-teaching-german/a-215435>, 2025.
- [18] Commission, "Erasmus+ Programme Guide," 2018.
- [19] Dulingo, "Duolingo for schools," <https://schools.duolingo.com/>.
- [20] B. e. al., "Higher Education Horizon Report.," 2019.



**Georgiana VOICA** graduated from the Faculty of Cybernetics, Statistics and Economic Informatics of the Academy of Economic Studies in 2023. She is currently pursuing a Master's degree in Business Support Database Systems. Her main areas of interest and work are focused on programming and working with SAP technologies. Throughout both bachelor's thesis and dissertation, she has concentrated on developing language learning applications that integrate advanced functionalities such as artificial intelligence (AI), optical character recognition (OCR) and speech recognition.

## Comparison of the Performance of SQL and NoSQL Databases in Modern Business Applications

Aniela BORCAN, Diana CAPTARI, Emanuela-Cristina CARP  
Bucharest University of Economic Studies  
Faculty of Cybernetics, Statistics and Economic Informatics  
Bucharest, Romania

[borcananiela20@stud.ase.ro](mailto:borcananiela20@stud.ase.ro), [captaridiana20@stud.ase.ro](mailto:captaridiana20@stud.ase.ro), [carpemanuela20@stud.ase.ro](mailto:carpemanuela20@stud.ase.ro)

*Databases represent a fundamental element of modern business applications, directly influencing performance, scalability, and data management. This article explores the differences between SQL and NoSQL databases, analyzing the advantages and disadvantages of each technology based on criteria of performance, consistency, and specific applications. Databases are essential structures for storing and managing information, having a major impact on the functioning of business applications. The choice between a relational database (SQL) and a non-relational database (NoSQL) depends on several factors, including the type of data being handled, scalability requirements, and the data access model.*

**Keywords:** SQL databases (Structured Query Language), NoSQL databases (Not Only SQL), relational databases, SQL vs NoSQL, NoSQL usage, NoSQL performance, data consistency, scalability.

### 1 Introduction

In the era of digitalization, the volume of data generated by companies has increased exponentially, necessitating the use of efficient solutions for storing and managing information. Databases have evolved significantly from traditional relational models to more flexible alternatives, such as NoSQL databases. This evolution has been driven by the need for companies to manage unstructured and semi-structured data, as well as to support high performance and scalability requirements [1].

The choice between a relational database (SQL) and a non-relational database (NoSQL) depends on several critical factors, including:

- Type of data used;
- The complexity of the queries;
- The need for strict consistency;
- The requirements for horizontal or vertical scalability.

SQL databases, based on the relational model, offer well-defined structures and support for ACID transactions

(Atomicity, Consistency, Isolation, Durability), being widely used in critical fields such as finance and healthcare [2]. On the other hand, NoSQL databases are designed to handle large volumes of distributed data and are used in scalable web applications, Big Data analytics, and recommendation systems.

### The fundamentals of databases SQL and NoSQL

**SQL databases:** Relational databases (SQL) are structured on a tabular model, using the SQL (Structured Query Language) query language. SQL is a language used for managing and processing data in a relational database. In a relational database, information is organized in the form of tables, each having rows and columns that represent the attributes of the data and the relationships between them. Through SQL instructions, users can store, update, delete, search, and extract data from the database. At the same time, SQL allows

for the administration and optimization of database performance.

The SQL query language is essential due to its universal use in various software applications. Data analysts and developers learn and adopt SQL due to its excellent compatibility with multiple programming languages. For example, SQL queries can be integrated into applications developed in programming languages such as Java, contributing to the creation of robust solutions for data processing within large-scale relational database systems like Oracle or MS SQL Server.

SQL is also an accessible language, considering that it uses familiar keywords in English, thus facilitating its learning. For example, the keyword "SELECT" is used to extract data from a database, "FROM" to specify the source table, "WHERE" to apply filtering conditions, and "ORDER BY" to organize the results in a specific order [3].

**NoSQL databases:** The term "NoSQL" refers to non-relational databases, which use a data storage format different from that of relational tables. However, NoSQL databases can be queried through APIs, using idiomatic languages, declarative structured query languages, or example-based query languages. For this reason, these databases are often referred to as "not only SQL."

NoSQL databases are widely used in real-time web applications and Big Data management, their main advantage being the ability to provide superior scalability and availability.

Additionally, NoSQL databases are preferred by developers due to their flexibility, being able to quickly adapt to the ever-changing requirements of projects. These databases allow for data storage in a more intuitive and understandable way, being closer to the format in which they are used in applications, which reduces the need for additional transformations for storage or retrieval through specific NoSQL APIs.

Moreover, NoSQL databases are optimized to fully leverage cloud infrastructure, ensuring continuous and efficient operation [4].

#### **When to choose a NoSQL database**

Considering the necessity for companies and organizations to innovate rapidly, the ability to remain flexible and support operations at any scale becomes essential. NoSQL databases offer flexible schemas and support a variety of data models, making them ideal solutions for developing applications that require managing large volumes of data and low latency or reduced response times, such as web applications for online gaming and e-commerce platforms [4].

#### **When not to choose a NoSQL database**

NoSQL databases are generally built on a denormalized data model, supporting applications that use a small number of tables (or containers) and do not rely on references to correlate data, but rather on embedded records (or documents). Many traditional back-office applications of companies, such as finance, accounting, and resource planning, rely on highly normalized data to prevent anomalies and data duplication. Generally, these applications are not suitable for using a NoSQL database.

Another important aspect of NoSQL databases is the complexity of queries. Although these databases work excellently for simple queries involving a single table, when the complexity of the queries increases, relational databases become a more suitable choice. NoSQL databases usually do not offer advanced functionalities for complex joins, subqueries, or nesting queries in a WHERE clause.

In certain cases, however, it is not necessary to exclusively choose between a relational database and a non-relational one. In many situations, companies opt for hybrid databases, which allow the simultaneous use of relational and non-relational data models. This approach offers greater flexibility in managing

different types of data, while ensuring read and write consistency without compromising performance.

### **How does a NoSQL database work?**

NoSQL databases use various data models to manage and manipulate information. These databases are designed for applications that process large amounts of data and require low latency, as well as flexible data models. This is achieved by relaxing the strict data consistency requirements characteristic of other types of databases. To illustrate this, let's consider a simple database model for managing books.

In a relational database, information about a book is usually divided into multiple tables (a process called "normalization"), and the relationships between them are established through primary and foreign key constraints. For example, a table named "Books" might contain columns such as ISBN, Title, and a foreign key referencing an "Authors" table that contains columns like Author Name and Author ID. The relational model is created to ensure referential integrity between tables, and the data is normalized to reduce redundancy and to be optimized for efficient storage.

In contrast, in a NoSQL database, a book record is usually stored as a JSON document. In this case, for each book, information such as ISBN, Title, Edition Number, Author's Name, and Author's ID is stored together as attributes of a single document. This model is optimized for more intuitive development and horizontal scalability, allowing for efficient data management in a more flexible manner [5].

### **Types of NoSQL databases**

NoSQL databases are used in situations where storing data in tables is not optimal. They use various storage formats, each with specific applications. There are six main types of NoSQL databases, each with distinct characteristics.

- **Key-value pair-based databases**

These databases allow for high separability and extensive horizontal scalability, which is impossible to achieve in a similar manner with other types of databases. They are ideal for applications that require high performance and low latency, such as gaming, advertising, and IoT applications. A notable example is Amazon DynamoDB, which offers consistent performance with a latency of a few milliseconds, even at large scale. These features were essential for migrating Snapchat Stories to DynamoDB, given the high volume of writes recorded.

- **Document-type databases**

In this category, data is stored in the form of documents in formats such as JSON, which are directly compatible with the objects used in the application's code. These databases are extremely flexible and allow for the storage and querying of data in an intuitive way, like how data is structured in the application's code. Document models are useful for applications that involve catalogs, user profiles, or content management systems. Examples of document databases are Amazon DocumentDB (compatible with MongoDB) and MongoDB, which offer intuitive APIs for agile development.

- **Graph databases**

Graph databases are used for applications that work with complex datasets, such as social networks, product recommendations, fraud detection, and knowledge graphs. These databases are optimized for storing and manipulating relationships between entities. Amazon Neptune is a fully managed service that supports Property Graph and Resource Description Framework (RDF) graph models. Other examples include Neo4j and Giraph.



- **In-memory databases**

These databases are used in applications that require quick responses in microseconds, such as gaming and online advertising, where traffic can spike suddenly. Amazon MemoryDB for Redis is an in-memory database service that reduces data read latency and ensures reliability at scale, making it ideal for microservices-based applications. Amazon ElastiCache, compatible with Redis and Memcached, serves workloads that require low latency and high throughput, while Amazon DynamoDB Accelerator (DAX) enhances its performance by reading data several times faster.

- **Databases for searches**

Many applications generate logs to assist with troubleshooting and problem analysis. Amazon OpenSearch is a service dedicated to the real-time analysis and visualization of data streams, indexing, aggregating, and searching information from logs and semi-structured data. OpenSearch also offers high-performance full-text search capabilities, being used, for example, by Expedia, which manages 30 TB of data and 30 billion documents for cost optimization and operational monitoring.

## SQL Data Types

Databases SQL uses a well-defined set of data types, which are applied in a structure of tables with rows and columns. Each column in a table is defined with a specific data type, and these types are very rigid and normalized.

- Numbers:

- INT: For storing integer numbers.
- DECIMAL or NUMERIC: For numbers with high precision decimals.
- FLOAT or DOUBLE: For floating-point numbers.

- Strings:

- VARCHAR: For variable-length character strings.
- CHAR: For fixed-length character strings.
- TEXT: For long character strings.

- Date and time:

- DATE: Used to store only the date (year, month, day).
- TIME: Used to store only the time.
- DATETIME or TIMESTAMP: Used to store both the date and time in a single field.

- Boolean:

- BOOLEAN: For storing values like true/false.

- Binary:

- BLOB: For binary data, such as images or files.

- Others:

- ENUM: To define a limited list of possible values for a field.
- SET: To store a set of values.

## Tools and Frameworks for Managing NoSQL Databases

**MongoDB Atlas** is a fully managed platform for MongoDB, which offers cloud services for managing MongoDB databases. It allows developers to create, monitor, and scale MongoDB-based applications without having to manage the infrastructure [6].

### Use cases:

- Development of modern web applications.
- Mobile and IoT applications.
- Content management systems and user profiles.

**Apache Cassandra** is a distributed NoSQL database that is extremely scalable and fault-tolerant. It is ideal for applications that require a large volume of distributed data and cannot tolerate service interruptions.

### Use cases:

- Applications that manipulate Big Data, such as real-time data analysis.

- Write-heavy distributed systems.
- Social media and e-commerce applications that require massive scalability.

**Redis** is an in-memory NoSQL database that offers high performance for storing and manipulating temporary data, such as sessions, queues, and caches. Redis is widely used in applications that require low latency and fast response.

**Use cases:**

- Storing sessions in web applications.
- Caching to accelerate data access.
- Message queues and jobs in distributed applications.

**Couchbase** is a document-oriented NoSQL database that combines the features of document databases with those of key-value databases. It is optimized for applications that require high scalability and performance.

**Use cases:**

- Web and mobile applications with high scalability and performance requirements.
- Applications for managing semi-structured data.
- Authentication and user management systems.

**Amazon DynamoDB** is a fully managed NoSQL database service offered by AWS, which ensures automatic scalability and high performance for applications that require a large volume of data and low latency [6].

**Use cases:**

- Mobile and web applications that require rapid scalability and low costs.
- IoT and Big Data applications.
- Management of structured and semi-structured data.

## Comparing performance between SQL and NoSQL

### Performance in CRUD operations:

CRUD operations (Create, Read, Update, Delete) are essential for any database management system. Their performance varies depending on the database architecture.

**SQL:** Read efficiency, write difficulties:

- SQL databases are optimized for complex queries and read operations. Due to advanced indexing and normalization, they allow for rapid data retrieval.
- Write operations (Create, Update) can be slower due to strict referential integrity checks and asynchronous index updates.

**NoSQL:** Optimized for writing, high flexibility:

- NoSQL databases are faster for write operations because they do not require strict integrity checks.
- They use techniques such as partitioning (sharding) and replication to efficiently distribute the workload.

Practical example:

Financial analysis applications: SQL is preferred due to complex queries.

Logging systems: NoSQL is ideal due to its high data write speed.

### Consistency Models and Transaction Management:

Data consistency is a critical factor in choosing a database, influenced by the ACID and BASE models.

**SQL:** The ACID Model (Atomicity, Consistency, Isolation, Durability):

- Offers reliable and secure transactions.
- It is essential in financial, medical, or other fields where data integrity is crucial.
- Cost: Performance can be affected by the bottlenecks caused by maintaining strict consistency.

**NoSQL:** The BASE Model (Basically Available, Soft-state, Eventual consistency):

- It offers high availability but allows eventual consistency.
- It is used in distributed applications, such as social networks or caching systems.
- Cost: Data can be temporarily inconsistent.

Practical example:

- Banks and hospitals: They require strict consistency, so SQL is preferred.
- Social networks: They can tolerate eventual consistency, so NoSQL is more suitable.

### Scalability and Availability:

Scalability is essential for managing the growth of data volume and the number of users.

SQL: Vertical scalability, limitations on horizontal:

- Performance improvement is achieved by adding hardware resources (vertical scalability).
- Horizontal scalability (adding nodes) is difficult due to the complex relationships between tables.

NoSQL: Horizontal scalability, efficient distribution:

- Allows for the easy addition of new nodes.
- Techniques such as sharding and replication ensure balanced data distribution.
- It is used in large systems, such as Amazon, Facebook, or Google.

Practical example:

- Banking systems: The vertical scalability of SQL is sufficient for many financial applications.
- E-commerce: NoSQL allows for rapid scaling and handling of high traffic [7].

### Current trends and the future of databases in modern business applications

In today's digital era, databases play an essential role in managing information

for modern business applications. Current applications require high scalability, optimized performance, and great flexibility to handle the volume of data in an ever-changing world. In this context, recent and future trends in the field of databases, both SQL and NoSQL, are directed towards solutions that meet the complex needs of modern businesses.

Another important trend in database development is the migration to cloud databases and serverless databases. Currently, many business applications are migrating to cloud-based solutions to reduce costs and the complexity of infrastructure management. Cloud databases, such as AWS Aurora or Google Cloud SQL, offer flexibility and scalability, being managed by cloud service providers. Additionally, the concept of serverless databases, exemplified by Firebase Firestore and AWS Aurora Serverless, allows applications to automatically scale based on traffic needs, without the need to manage a server or fixed resources. These solutions are extremely attractive for startups and businesses that do not want to invest heavily in infrastructure but need performance and scalability as their business grows.

At the same time, another significant trend in database usage is the adoption of graph databases, especially for analyzing complex data. Graph databases, such as Neo4j or Amazon Neptune, are increasingly used by businesses that need to analyze complex relationships between entities in their data systems, such as users, products, transactions, or connections in social networks. These databases allow for a much more intuitive representation of relationships between data, facilitating analyses and the discovery of relevant information, such as fraud detection, product recommendations, or supply chain optimization.

In addition to these, Artificial Intelligence (AI) and Machine Learning

(ML) are having an increasingly significant impact on the management and optimization of databases. AI algorithms are used to analyze query behavior and optimize database performance by dynamically adjusting allocated resources (storage, CPU, memory) based on application requirements. Additionally, AI is used to identify anomalies and predict potential system failures before they impact operations, thereby contributing to reduced downtime and increased reliability.

Another important direction for the future of databases in business applications is the integration of multi-model databases. Businesses that have diverse requirements regarding data types (relational, graph, document, etc.) can adopt database solutions that integrate multiple data models into a single platform. This allows them to meet various storage and processing requirements, providing greater flexibility in data management.

One of the most notable aspects of the evolution of databases in modern business applications is the transition to solutions capable of providing horizontal scalability. While traditional SQL databases were excellent for managing complex relationships between data, they had limitations in terms of horizontal scalability, meaning they could not efficiently scale as the volume of data grew exponentially. In contrast, NoSQL databases, such as Cassandra and MongoDB, are built to support excellent horizontal scalability, allowing business applications to manage large volumes of data distributed across multiple servers. These solutions are ideal for applications that handle semi-structured or unstructured data, such as those in the field of IoT (Internet of Things) or Big Data, making them a popular choice for businesses that need to process data quickly in a distributed manner [8, 9].

### **Case study: Migration from SQL to NoSQL**

A notable example of migration from SQL to NoSQL is the case of Netflix, a global video streaming platform, which migrated from a relational database (SQL) architecture to an infrastructure based on NoSQL databases, such as Cassandra and Amazon DynamoDB. The decision was motivated by the need to scale rapidly and efficiently as their storage and data processing requirements grew exponentially, in the context of an increasing number of users and video content.

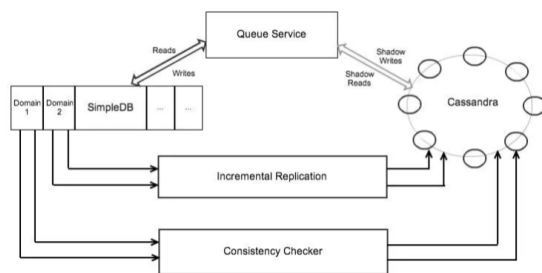
**Data Migration:** The migration was a multi-step process, involving an initial massive migration ("forklift"), followed by incremental replication and data consistency validation. These stages aimed to transfer data from SimpleDB to Cassandra without interrupting the service and without degrading performance:

- **Forklift (Initial Migration):** In the first stage, all data from SimpleDB was copied to Cassandra. The process was carried out over the course of 30 hours, and the number of threads and instances was adjusted to minimize the impact on SimpleDB's performance during the migration.
- **Incremental Replication:** After the forklift, changes from the Queue continued to be written only in SimpleDB. Incremental replication ensured that the changes made to the data in the post-migration period were also reflected in Cassandra.
- **Consistency Checker:** To verify if the migration and incremental replication were correct, a consistency checker was used. It compared the data between SimpleDB and Cassandra, updating only the records that did not match between the two databases.
- **Shadow Writes:** At this point, data continued to be read from SimpleDB,

but for each Queue update, the data was also written to Cassandra. This allowed for the validation of performance and consistency of writes in Cassandra, without disrupting the existing user flow.

- **Shadow Reads (Parallel Reads):** After shadow writes were implemented, the shadow reads functionality was also activated, where the data from the Queue was read from both SimpleDB and Cassandra. The differences between the data from the two sources were monitored and corrected.
- **Ending the use of SimpleDB:** After the discrepancies between SimpleDB and Cassandra were minimal (<0.01%), SimpleDB was eliminated, and Cassandra became the primary source of truth for the data in Queue. All reads and writes to Queue were now exclusively managed by Cassandra.

**Results:** The migration from SimpleDB to Cassandra allowed Netflix to improve the scalability, performance, and availability of the Queue service, without interruptions and with minimal impact on users.



**Fig. 2.** Migration from SQL to NoSQL

#### **Advantages of migration:**

- **Horizontal scalability:** NoSQL allows the addition of nodes to manage large volumes of data, unlike SQL, which has limitations in this regard.
- **Improved write performance:** NoSQL is more efficient for write operations, ideal for high-traffic platforms like Netflix.

- **Flexibility:** NoSQL allows for the storage of semi-structured and unstructured data, which is suitable for movie metadata and user interactions.
- **Availability and fault tolerance:** NoSQL, such as Cassandra, offers continuous availability even in the event of hardware failures.

#### **Disadvantages of migration:**

- **Costs and complexity in migration:** Migrating from SQL to NoSQL required significant investments in resources and time, and development teams had to learn new paradigms.
- **Lack of ACID transactions:** NoSQL does not guarantee the same complex transactions and data integrity as SQL, which can be problematic for financial applications.
- **Challenges with complex queries:** NoSQL databases are not ideal for complex queries, such as joins between tables, requiring alternative solutions like Apache Kafka and ElasticSearch.
- **Eventual consistency:** NoSQL uses the eventual consistency model, which can lead to temporary data inconsistencies.

#### **Migration issues and solutions:**

- **Team adaptability:** Teams had to learn the new architectures and paradigms. The solution was to form specialized teams in NoSQL and conduct training sessions.
- **Data inconsistency during migration:** A gradual migration process was used, with both databases running in parallel, to minimize risks.
- **Monitoring:** An internal monitoring system was implemented to track the performance and health of the distributed NoSQL database [10, 11, 12].

**What are hybrid solutions and interoperability between SQL and NoSQL?**

Hybrid solutions allow the integration of both relational and non-relational databases into a single technological architecture, to meet a wide range of requirements related to performance, scalability, flexibility, and data consistency. These solutions aim to utilize the best features of each type of database according to the specific needs of the applications, without sacrificing performance or data integrity.

#### **Challenges of integration:**

- **Implementation complexity:** Integrating SQL and NoSQL databases into a hybrid solution can be technically complex. This can involve managing multiple data storage solutions and query mechanisms, which can lead to increased operational costs and the need for specialized technical teams.
- **Data consistency:** In some cases, NoSQL databases allow for a weaker consistency model (in accordance with the eventual consistency model), while SQL databases guarantee consistency through ACID transactions. In a hybrid system, it can be difficult to maintain uniform consistency between the two databases, which can lead to issues related to data integrity.
- **Integrating queries and APIs:** Another challenging aspect is the integration of queries between SQL and NoSQL databases. Each type of database has its own query language (SQL for SQL and custom APIs for NoSQL), and combining them may require additional development effort.

#### **Examples of hybrid systems that combine SQL and NoSQL:**

- **Microsoft Azure Cosmos DB:** A notable example of a hybrid solution is Microsoft Azure Cosmos DB, which supports both SQL and NoSQL data models. Cosmos DB offers the integration of a relational database model

based on SQL, as well as the possibility to use various NoSQL models, such as document, graph, and key-value models. It thus offers a scalable platform that can manage both structured and semi-structured or unstructured data as applications develop [13].

- **MongoDB with SQL integration:** Although MongoDB is a document-based NoSQL database, there are solutions like MongoDB Atlas that allow integration with SQL databases. This allows developers to manage both relational and semi-structured data simultaneously. Additionally, MongoDB includes advanced search and indexing functionalities that are compatible with SQL queries, and users can also use APIs to interact with the data in a hybrid manner [14].
- **PostgreSQL with NoSQL extensions:** PostgreSQL, one of the most popular SQL relational databases, has implemented extensions to support NoSQL functionalities. For example, with the HStore extension, PostgreSQL can store semi-structured data in a key-value pair format. Additionally, with the JSONB extension, PostgreSQL can store and query JSON documents, making it compatible with the data models used by NoSQL databases [15].
- **Amazon Aurora and DynamoDB:** In the AWS ecosystem, Amazon Aurora (a relational database compatible with MySQL and PostgreSQL) can be used together with Amazon DynamoDB, a NoSQL database. Aurora is used for managing transactions and relational data, while DynamoDB is used for managing data that

requires scalability and low latency. This combination offers a robust hybrid solution that can be rapidly scaled according to the application's needs [16].

- **Couchbase and SQL:** Couchbase is a document-oriented NoSQL database that provides support for SQL queries through the N1QL language. This allows users to write standard SQL queries to query data stored in JSON format, thus combining the advantages of a NoSQL model with the familiarity of SQL queries [17].

### **Modern database architectures: Polyglot Persistence and microservices**

In modern business applications, database performance is closely tied to how they are integrated into the overall system architecture. The microservices model, increasingly adopted, assumes that each functional module can use a database specific to its needs. This principle is called Polyglot Persistence and allows the combination of relational databases (SQL) with non-relational ones (NoSQL) [18].

A practical example is in e-commerce: the SQL database is used for processing orders and transactions where consistency and integrity are critical while NoSQL is preferred for comments, reviews, or recently viewed products, where flexibility and speed are essential. This approach brings increased flexibility and allows for independent scaling of components. At the same time, technologies like Apache Kafka facilitate data exchange between microservices, eliminating bottlenecks between components and providing a more adaptable and stable architecture.

According to Fowler [18], using a combination of databases within a complex ecosystem helps avoid the limitations specific to each type of technology, thereby increasing the application's reliability.

### **Artificial Intelligence and Machine Learning in databases**

The performance of modern databases no longer depends solely on indexes or well-designed structures, but also on the use of advanced technologies such as artificial intelligence (AI). More and more solutions use machine learning algorithms to automatically optimize how queries are executed or resources are managed [19].

For example, Oracle Autonomous Database can automatically adjust the execution plan of a query or decide when and how to create indexes, without human intervention. On the NoSQL side, systems are becoming increasingly intelligent in resource management: they learn application behaviors and pre-load the most frequently accessed data, optimizing cache memory and response times.

This AI-based automation not only optimizes performance but also helps reduce human errors. In a large-scale organizational environment, where data is critical and updates are frequent, AI becomes a reliable ally for ensuring service continuity [20]. Özsu and Valduriez explain in detail this type of AI integration in distributed systems in their seminal work [19].

### **Real-time data management**

We live in an era of instantaneity, users and businesses expect real-time reactions and analyses. In this context, databases must be capable of processing large volumes of data continuously and quickly [21].

NoSQL solutions are often the primary choice for such scenarios. They easily integrate with real-time processing platforms like Apache Kafka or Apache Flink and can handle real-time data streams, such as bank transactions, messages from IoT sensors, or user activities on a website.

On the other hand, SQL has also evolved, and some modern solutions (such as



Snowflake or ClickHouse) offer very good performance in the near-instant processing of semi-structured data. However, these usually come with a higher cost and a steeper learning curve. As Tyler Akidau shows in the work "Streaming Systems" [21], systems capable of handling real-time data are essential for modern applications that cannot afford processing delays.

### **Cost comparison: SQL vs NoSQL in the cloud**

Beyond performance, cost is one of the most sensitive criteria for any company. In the cloud, the differences between SQL and NoSQL become evident from the perspective of pricing and flexibility. SQL is usually billed based on reserved resources (processor, RAM, storage space), which means you pay even when the database is inactive. In contrast, many NoSQL services such as DynamoDB, allow flexible billing based on the actual number of operations performed [22]. NoSQL also has the advantage of open-source implementations that can completely eliminate licensing costs. However, in such cases, other hidden costs come into play: maintenance teams, security, backup. It is important for a Total Cost of Ownership (TCO) analysis to be conducted before the final decision is made [23].

According to Joe Weinman, in "Cloudonomics" [23], a balanced approach between performance and costs is essential to achieve a sustainable competitive advantage in dynamic digital environments.

### **The impact of database choice on user experience**

The choice of database not only affects the technical aspect of the application but also the way users interact with it. If the data loads slowly, if there are consistency errors, or if certain functions are unavailable at key moments, the user will perceive the application as weak,

regardless of how well the code is written.

NoSQL databases, by their distributed nature and optimization for fast reads, can provide an extra level of fluidity. This is especially evident in social applications (e.g., Facebook with TAO) or messaging apps (e.g., Snapchat with DynamoDB), where responses need to come in milliseconds.

In contrast, applications where accuracy is more important than speed, such as financial or medical reporting will benefit more from an SQL database, which offers strict consistency and secure transactions [24].

### **Performance benchmarks: SQL vs NoSQL in practice**

To truly understand the performance differences between SQL and NoSQL, it is important to analyze concrete data from independent benchmarks. For example, in the tests conducted by the Yahoo! Cloud Serving Benchmark (YCSB), it was found that MongoDB (NoSQL) has superior performance in distributed write operations, while PostgreSQL (SQL) stands out for its efficiency in complex JOIN queries [25, 26].

These benchmarks highlight that the choice of database is not universal: for applications involving complex aggregations and multiple relationships between data, SQL is superior. In contrast, for applications that require fast writes and horizontal scalability, NoSQL is more efficient. The study conducted by Zdravovski et al. (2023) [26] confirms this conclusion, indicating that MySQL offers stable performance in OLTP tasks, while MongoDB excels in handling semi-structured data in Big Data tasks.

### **Choosing the database based on the domain**

The selection between SQL and NoSQL databases must be guided by the specific requirements of the application domain.

For example, in the financial sector, where data consistency and integrity are essential, SQL databases are often preferred due to their compliance with ACID properties [27]. In contrast, in fields such as social networks or video game applications, where scalability and flexibility are crucial for managing large volumes of data and high interaction rates, NoSQL databases offer significant advantages.

A notable example is Netflix's use of Cassandra for managing personalized recommendations, demonstrating NoSQL's ability to quickly process large amounts of unstructured data and provide real-time responses. In e-commerce, a hybrid approach is often preferred: SQL is used for transactions, while NoSQL is used for the product catalog and user interactions.

### **Challenges in adopting NoSQL**

Although NoSQL databases offer numerous benefits, their adoption also comes with significant challenges. One of the main difficulties is the lack of native support for ACID transactions, which can affect data consistency in critical applications [28]. To address this issue, researchers such as Alflahi et al. (2023) have proposed transaction management models in MongoDB that include specialized locking algorithms and logical separation of read/write operations, improving both performance and data consistency [28].

Another challenge is the lack of standardization among different NoSQL solutions, which can lead to interoperability issues and risks related to vendor lock-in. Therefore, it is essential for organizations to conduct pilot tests and choose vendors with active support and a strong technical community.

### **NewSQL and multi-model databases**

To combine the advantages of relational databases with the flexibility and scalability offered by NoSQL, hybrid solutions such as NewSQL and multi-

model databases have emerged. NewSQL aims to provide the performance and scalability of NoSQL while maintaining the ACID properties characteristic of SQL [29]. These systems are designed to efficiently manage online transaction processing (OLTP) tasks in the context of Big Data.

On the other hand, multi-model databases, such as ArangoDB and OrientDB, allow for the storage and querying of data using multiple models (documents, graphs, columns), providing increased flexibility in managing various types of data and complex relationships [27]. These solutions represent a bridge between the advantages of both worlds, allowing developers to choose the most suitable data model for each component of their application.

### **Consistency models in distributed systems: trade-offs and options**

In modern distributed architectures, the choice of consistency model has a significant impact on the performance and behavior of the application. NoSQL databases like Cassandra or DynamoDB adopt the eventual consistency model, which offers availability and scalability at the expense of immediate consistency. On the other hand, traditional SQL systems promote strong consistency, ideal for critical applications where data must always be accurate [30].

More and more modern databases offer configurable consistency options. For example, Cosmos DB allows the choice between five consistency levels (including bounded staleness and session consistency), giving developers fine control over the balance between performance and accuracy [31]. This flexibility comes with direct implications for architectural design: stricter consistency involves increased latency and higher replication costs, while weak consistency can lead to users viewing outdated data.

Another important aspect is the integration of consistency models based on the type of data and usage. In critical applications, such as banking or healthcare systems, strict consistency is preferred to ensure the absolute correctness of transactions. In contrast, for social network applications, a small temporary lag between the states of replicas is acceptable, as long as availability and scalability are maximized.

The CAP theorem (Consistency, Availability, Partition tolerance) remains a fundamental theoretical benchmark in this context. In a distributed environment, only two of the three characteristics can be guaranteed simultaneously. The choice of the right strategy thus depends on the application's priorities: performance, reliability, or absolute consistency.

### **Query and index optimization in dynamic environments**

Query optimization is a constant challenge in managing large and dynamic databases. Traditional query optimization algorithms rely on estimated costs and the analysis of data statistics, but in modern environments characterized by constantly changing data, this approach is not always sufficient.

PostgreSQL, for example, uses a cost-based planner that evaluates all possible data access strategies and selects the one with the lowest estimated cost. However, in distributed environments, costs can vary over time, and dynamic query reoptimization becomes a necessity.

NoSQL databases, such as MongoDB and Couchbase, offer powerful tools for performance optimization through compound indexes, geospatial indexes, or TTL indexing. For example, in location-based applications, geospatial indexes dramatically reduce the response time for queries involving coordinates.

Another growing trend is the use of "adaptive indexing" algorithms, which dynamically adjust the structure of the

indexes based on user access patterns. This technique is effective in applications with variable traffic, as it eliminates the need for constant manual reconfiguration. Modern observability and profiling tools, such as Query Profiler in MongoDB or EXPLAIN ANALYZE in PostgreSQL, provide granular visibility into the time spent at each stage of execution. Thus, developers can quickly identify bottlenecks and propose targeted or strategic optimizations [32, 33].

### **Database security in modern architectures**

Database security is a constantly evolving field, considering the increase in the number and complexity of cyberattacks. In modern cloud-native architectures, data circulates through multiple layers and systems, significantly increasing the risk of interception, loss, or unauthorized modification.

In the case of relational databases, protection mechanisms are well-defined and include column or table-level encryption, role-based access control (RBAC) policies, and detailed operation logging. Systems like Oracle or SQL Server offer advanced auditing functionalities, including for preventing internal threats.

NoSQL databases have also evolved significantly in this regard. MongoDB, Couchbase, or Cassandra now offer support for full encryption (end-to-end), certificate-based authentication, and integration with complex identity management systems (LDAP, Kerberos). The new versions also include audit functions, protection against injection attacks, as well as field-level security controls.

Additionally, distributed systems require the protection of communication channels between nodes. The TLS 1.3 protocol is used for traffic encryption, and network access control is achieved through security groups, isolated subnets, and declarative firewall policies.

A key element is also context-based rights management. Modern systems implement models such as Attribute-Based Access Control (ABAC), which allow flexible policies based on user, device, or location attributes. Thus, a user can access a specific database only under certain conditions (e.g., only within the company's network and during working hours).

In the future, a closer integration between databases and cybersecurity threat analysis platforms is expected. These will enable the automatic identification and blocking of suspicious behavior in real-time, through behavioral analysis and machine learning [34, 35, 36].

### **Integration of databases into Big Data and Artificial Intelligence ecosystems**

An increasingly important development direction in modern business applications is the integration of databases with Big Data processing platforms and artificial intelligence, aspects also addressed in the recent works of Abadi and Pavlo [37, 38]. In this context, both SQL and NoSQL play essential roles.

SQL databases are used in the stages of data preprocessing and cleaning, where the rigid structure and transactionality ensure data integrity. On the other hand, NoSQL databases are preferred for storing and quickly accessing large volumes of unstructured data, used in training machine learning algorithms.

For example, Apache Spark offers native connectors for both PostgreSQL and MongoDB or Cassandra, allowing for distributed in-memory processing. Thus, data can be extracted from relational databases, transformed, and then combined with NoSQL sources for predictive analytics or real-time recommendation systems.

More and more modern databases are starting to integrate artificial intelligence-based functionalities directly into their structure. Oracle Autonomous Database and Microsoft SQL Server use machine

learning algorithms for auto-tuning and anomaly detection, according to the technical documentation and research published by Microsoft and Oracle [39] for automatic query optimization, anomaly detection, and index proposal.

Integration with artificial intelligence also entails challenges: massive volumes of data must be managed, processing ethics ensured, and traceability of automated decisions guaranteed. For this reason, hybrid systems that combine the flexibility of NoSQL with the stability offered by SQL become ideal solutions for complex organizations with diverse data management needs.

### **Conclusions**

The choice between SQL and NoSQL depends on the application's context. There is no universal solution. Relational databases are essential for critical applications that require complex transactions and strict consistency, while NoSQL offers flexibility and scalability for modern applications that handle large volumes of unstructured data.

Migrating to NoSQL brings benefits, but also challenges. Cases like Netflix highlight the advantages of horizontal scalability and superior performance for distributed data. However, the transition from SQL to NoSQL involves significant technical challenges, particularly in maintaining data consistency and adapting the architecture.

Hybrid and multi-model architectures are increasingly being adopted. Combining the flexibility of NoSQL with the reliability of SQL, along with approaches like Polyglot Persistence and microservices, offers an adaptable solution for complex organizations, allowing for independent scaling and varied technological integration.

Artificial intelligence fundamentally changes the way databases are designed and managed. AI integration allows for automatic query optimization, dynamic resource adjustment, and early error

detection, reducing human intervention and increasing operational efficiency.

AI and machine learning are becoming integral to the future direction of databases. Machine learning algorithms are increasingly used for predictive analysis, intelligent resource allocation, and anomaly detection, while graph databases are gaining ground in analyzing complex relationships and information networks.

Cloud-native technologies and serverless architectures are redefining database infrastructure. Migrating to cloud-based solutions and pay-per-use models allows companies to reduce operational costs and dynamically manage resources without compromising performance or security.

## References

- [1] "The Evolution of Database Technologies", *IEEE Journals*, 2023
- [2] "SQL vs NoSQL: Choosing the Right DBMS," SimeonOnSecurity. [Online]. Available: <https://ro.simeononsecurity.com/articles/sql-vs-nosql-choosing-the-right-database-management-system>. [Accessed: March 24, 2025]
- [3] "Ce este SQL?" *NewTech*. [Online]. Available: <https://www.newtech.ro/blog-ce-este-sql/>. [Accessed: March 24, 2025].
- [4] Oracle, "What is NoSQL?" [Online]. Available: <https://www.oracle.com/ro/database/nosql/what-is-nosql/>. [Accessed: March 24, 2025].
- [5] CodeGym, "Hibernate: SQL vs NoSQL." [Online]. Available: <https://codegym.cc/ro/quests/lectures/ro.questhibernate.level19.lecture00>. [Accessed: March 24, 2025].
- [6] Guru99, "SQL Tools Overview." [Online]. Available: <https://www.guru99.com/ro/sql-tools.html>. [Accessed: March 25, 2025].
- [7] "Conf-TehStiint-UTM-StudMastDoct-2024-V1-p585-588.pdf," *Proceedings of UTM Conference*, pp. 585–588, Chisinau, Moldova, 2024. [Online]. Available: <https://repository.utm.md/bitstream/handle/5014/27928/Conf-TehStiint-UTM-StudMastDoct-2024-V1-p585-588.pdf>. [Accessed: March 25, 2025].
- [8] Budibase, "Data Management Trends." [Online]. Available: <https://budibase.com/blog/data/data-management-trends/>. [Accessed: March 24, 2025].
- [9] "Database Management Trends in 2024," *Dataversity*. [Online]. Available: <https://www.dataversity.net/database-management-trends-in-2024/>. [Accessed: March 24, 2025].
- [10] Netflix Tech Blog, "Netflix Queue Data Migration for a High-Volume Web Application." [Online]. Available: <https://netflixtechblog.com/netflix-queue-data-migration-for-a-high-volume-web-application-76cb64272198>. [Accessed: March 26, 2025].
- [11] Netflix Tech Blog, "NoSQL at Netflix." [Online]. Available: <https://netflixtechblog.com/nosql-at-netflix-e937b660b4c>. [Accessed: March 26, 2025].
- [12] P. Krill, "Big Movies, Big Data: Netflix Embraces NoSQL in the Cloud," *InfoWorld*, 2012. [Online]. Available: <https://www.infoworld.com/article/2171162/big-movies-big-data-netflix-embraces-nosql-in-the-cloud.html>. [Accessed: March 26, 2025].
- [13] Microsoft, "Azure Cosmos DB Documentation." [Online]. Available: <https://learn.microsoft.com/en-us/azure/cosmos-db/>. [Accessed: March 26, 2025].
- [14] MongoDB, "Atlas Database Platform." [Online]. Available: <https://www.mongodb.com/products/>

- [platform/atlas-database](#). [Accessed: March 26, 2025].
- [15] PostgreSQL, "JSON Data Types." [Online]. Available: <https://www.postgresql.org/docs/current/datatype-json.html>. [Accessed: March 26, 2025].
- [16] Amazon Web Services, "AWS Home." [Online]. Available: <https://aws.amazon.com/>. [Accessed: March 27, 2025].
- [17] Couchbase, "Documentation Home." [Online]. Available: <https://docs.couchbase.com/home/index.html>. [Accessed: March 27, 2025].
- [18] M. Fowler, "NoSQL and Polyglot Persistence," martinfowler.com, 2012.
- [19] T. Özsu, P. Valduriez, *Principles of Distributed Database Systems*, Springer, 2020.
- [20] Oracle, "AI-powered database optimization," oracle.com, 2023.
- [21] T. Akidau et al., *Streaming Systems: The What, Where, When, and How of Large-Scale Data Processing*, O'Reilly Media, 2018.
- [22] Amazon Web Services, "Amazon DynamoDB Pricing." [Online]. Available: <https://aws.amazon.com/dynamodb/pricing/> [Accessed: March 27, 2025].
- [23] J. Weinman, *Cloudonomics: The Business Value of Cloud Computing*, Wiley, 2012.
- [24] M. Stonebraker, "The Case for NewSQL Databases" *Communications of the ACM*, 2018.
- [25] Yahoo! Cloud Serving Benchmark, "YCSB Project" GitHub Repository, 2023.
- [26] E. Zdravevski et al., "Comparison of SQL and NoSQL databases with different workloads: MongoDB vs MySQL evaluation," *International Journal of Database Management Systems*, vol. 6, no. 3, pp. 1-14, 2023.
- [27] S. Patro, "SQL vs. NoSQL: Choosing the Right Database for Your Data Needs," *LinkedIn*, 2024.
- [28] A. A. E. Alflahi et al., "An Enhanced Model for Transactional Consistency in MongoDB," *arXiv preprint arXiv:2308.13921*, 2023.
- [29] A. B. M. Moniruzzaman, "NewSQL: Towards Next-Generation Scalable RDBMS for Online Transaction Processing (OLTP) for Big Data Management," *arXiv preprint arXiv:1411.7343*, 2014.
- [30] E. A. Brewer, *CAP twelve years later: How the 'rules' have changed*, Computer, 2012.
- [31] Microsoft Azure Cosmos DB, *Consistency levels documentation*, 2024.
- [32] A. Pavlo et al., *Self-tuning database systems: a review*, *ACM Computing Surveys*, 2021.
- [33] M. Stonebraker et al., *Readings in Database Systems*, 5th Edition, MIT Press, 2015.
- [34] A. Gupta et al., *Security in Modern Databases: Challenges and Techniques*, *ACM Computing*.
- [35] S. Ghazali et al., *A Survey on Secure NoSQL Databases in Cloud Environment*, *Journal of Network and Computer Applications*, 2020.
- [36] Microsoft Docs, *SQL Server Security Features*, Microsoft.com, 2024.
- [37] D. Abadi, *Data Management in the Cloud: Limitations and Opportunities*, *IEEE Data Eng. Bull.*, 2020.
- [38] A. Pavlo, *Big Data meets Big Query: Lessons from Real-World Systems*, *Proceedings of the VLDB Endowment*, 2021.
- [39] Oracle & Microsoft, *AI in Databases: Technical Briefs and Innovations*, Whitepapers, 2023.



**Aniela BORCAN** graduated in 2023 with a Bachelor's degree in Economic Cybernetics from the Faculty of Cybernetics, Statistics and Economic Informatics at the Bucharest University of Economic Studies. She continued her studies at the same institution, earning a Master's degree in Databases – Support for Business in 2025. With a background in both IT and the financial industry, Aniela has contributed to the development of digital solutions aimed at enhancing business efficiency and enabling data-driven decision-making. She has practical experience in backend development and database management, with key interests in scalable data infrastructures, process automation and the integration of secure and efficient digital banking systems.



**Diana CAPTARI** completed her undergraduate studies in 2023, earning a degree in Economic Cybernetics from the Faculty of Cybernetics, Statistics and Economic Informatics at the Bucharest University of Economic Studies. She further pursued graduate studies at the same university, obtaining a Master's degree in Data Bases – Support for Business in 2025. Currently active in the banking industry, she works as a programmer, with hands-on experience in software development and database technologies tailored to financial services. Her professional interests focus on digital solutions in banking, optimization of data systems, and the integration of programming tools in business processes.



**Emanuela-Cristina CARP** graduated in 2023 with a Bachelor's degree in Economic Cybernetics from the Faculty of Cybernetics, Statistics and Economic Informatics of the Bucharest University of Economic Studies. She obtained her Master's degree in *Data Bases – Support for Business* from the same institution in 2025. She is currently employed in the private sector, with professional experience in databases, artificial intelligence, and software development. Her research interests include database systems, intelligent data processing, and the integration of AI technologies into modern business solutions.



## A Taxonomy for Learning Cybersecurity with Artificial Intelligence

Mihai-Nicolae DULGHERU<sup>1</sup>, Andrei-Cătălin NICA<sup>2</sup>, Cătălin-Alexandru TRANDAFIRIDIS<sup>3</sup>

<sup>1</sup>The Bucharest University of Economic Studies

<sup>2</sup>The Bucharest University of Economic Studies

<sup>3</sup>National University of Science and Technology POLITEHNICA Bucharest

[dulgherumihai19@stud.ase.ro](mailto:dulgherumihai19@stud.ase.ro), [nicacatalin19@stud.ase.ro](mailto:nicacatalin19@stud.ase.ro), [alexandru@trandafiridis.ro](mailto:alexandru@trandafiridis.ro)

*The article explores the area of adaptive learning and proposes a taxonomy for supporting STEM teaching in schools and universities. The approach is based on the hypothesis that AI can improve the learning process by mapping the curriculum on a taxonomy that is used to create an individual knowledge map for each learner. The authors provide a detailed proposal of such a taxonomy for the area of cybersecurity at an undergraduate and master's level. It is based on a proposed EU taxonomy for cybersecurity.*

**Keywords:** *STEM, personalized learning, artificial intelligence, automated feedback*

### 1 Introduction

In these modern times, there is a very fast evolution of digital technologies. These technologies are starting to have components that are more and more developed using artificial intelligence. This is one of the reasons why education also needs to be updated regarding the methods of teaching and the technologies used. Also, taking into consideration the level of automation that can be seen everywhere, a new need for learning subjects like cybersecurity starts to grow. The progress that has been made in the past few years has opened new opportunities for the development of adaptive learning systems. These systems have a major importance in the efficiency analysis because they can analyze student performance in real time, generate personalized feedback, and improve student engagement, besides reducing the teacher's responsibilities.

The concept of adaptive learning in cybersecurity education needs a multilayered taxonomy. This taxonomy would work in a beneficial way for anyone who uses it. It would organize the content in a better manner, and it would adapt to the needs/ level of the student. The taxonomy proposed in the study is based on three main pillars: NIST CSRC taxonomy, ACM Classification System,

and the cognitive learning models (e.g., Bloom and Anderson). The first pillar is supposed to help by offering a robust thematic structure in the industry standards, while the second pillar is supposed to help with the academic rigor and proper coverage on this topic.

By adding these three pillars into a single framework, we are going to create a tool for an infrastructure system that is able to deliver personalized instructions, be scalable for automating assessments, tracking learner progress, and ensuring compliance with international regulations (e.g., GDPR or NIS Directives).

Taking all of these into consideration, this taxonomy is going to act as a bridge between technical proficiency and pedagogical information, helping both students and teachers and making them both more engaged in the process of education.

### 2 Literature review

This section contains a summary regarding the state-of-the-art in adaptive learning. We have identified a series of representative articles and a relevant taxonomy supported at the EU level.

An extremely complex task is assessing student knowledge in higher education. This requires the combination of traditional methods and technology-enhanced tools. The former are represented by projects,

portfolios, and presentations. The latter refers to e-portfolios, simulations, and automated feedback systems used to achieve optimal outcomes. Combining these two categories of methods opens new pathways for assessment, interactivity, and adaptability. By aligning assessment tools with learning objectives and understanding the challenges associated with each method, educators can gain a comprehensive view of students' competencies [1].

Student assessment is important both individually, for students to understand their progress, and collectively, to obtain a clear overview of the group's current knowledge status. It is often impossible to implement an effective feedback system that genuinely benefits students. In the academic environment, this is the result of a large number of students per professor and limited resources, especially time. Additionally, occurrences of human error are to be expected when student work assessments are performed by human evaluators. This results in students forming a mistaken perception of feedback, which leads them to perceive it as a burden or judgment rather than an opportunity to improve their skills.

In the case of classic feedback provided via online platforms, all the answers given to students are identical and lack personalization. Such feedback fails to pinpoint the exact mistake made by a student, hindering their ability to understand precisely what went wrong or how they might avoid the mistake in the future. Typically, an expert in the field must provide these general answers.

In contrast, feedback generated by an LLM is automatically produced based on students' answers in real time. It specifically addresses each student's questions and identifies their mistakes. Additionally, it offers precise guidance on avoiding similar errors without expert intervention. This method can also be combined with cognitive taxonomies.

This results in providing tailored feedback at various cognitive levels.

Furthermore, metrics generated from LLM feedback are personalized. These enable the lesson content to be adapted based on student performance [2].

LMS platforms can effectively support students' cognitive progress through explanatory and formative feedback. These provide clear explanations, often accompanied by additional helpful resources. The simple feedback merely states whether an answer is correct or incorrect. In contrast, explanatory feedback actively engages students in the learning process, clarifies concepts, and uses specific, real-time examples. These are generated by LLMs to help students comprehend their mistakes clearly and precisely.

The real-time feedback provided to evaluated students can greatly assist the platform in identifying concepts that students have not yet mastered, enabling adaptation of lesson content according to each student's needs [3].

Utilizing artificial intelligence-generated feedback can be significantly more relevant, timely, and efficient for students. Besides the fact that students who repeatedly review their knowledge tend to achieve better results compared to those who review only once, AI-generated feedback also relieves human evaluators of extra effort, saving a considerable amount of their time [4].

Adaptive learning systems have recently become increasingly popular due to their ability to customize the learning experience to the individual needs of each student by utilizing advanced algorithms for data analysis and performance monitoring. Previous research has demonstrated that artificial intelligence-based adaptive learning systems significantly improve students' educational outcomes. An example of this is provided by the article (Sari et al., 2024), which conducts a study involving 300 students and 50 teachers across primary to higher education, with a varied demographic structure. This study highlights that adaptive learning systems such as Smart

Sparrow and IBM Watson Education led to an increase in average assessment scores from 68.4 to 82.7 and improved course completion rates [5].

It is an educational method that is based on technology to personalize the learning process of students. This method, together with artificial intelligence and the support of machine learning algorithms, adjusts the teaching methods and the content of the lessons that are displayed to the student within the platform, to offer him an educational experience that helps him understand much more easily all the concepts for which the feedback obtained from the evaluation was not positive.

The Adaptive Learning method is based on several information accumulated during the evaluation by students. Among the information is information such as the current level of the student, his behavior at the time he receives feedback, or monitoring the answers to be able to observe any progress [6].

Knowledge can be modeled through a taxonomy, meaning hierarchical sets of models used to classify learning objectives according to their complexity level. Over time, multiple taxonomies have been developed and implemented to best accommodate various educational contexts. Due to their widespread popularity thus far, the taxonomies proposed by Bloom, Anderson, and Wilson have significantly impacted the academic world. However, due to the complexity of human beings and the human brain, there is a need to propose a new holistic taxonomy model of human thinking, which may necessitate further exploration in this field [7].

This hierarchical classification of learning objectives guides curriculum design and helps teachers formulate clear objectives that allow for the assessment of students at the cognitive level.

The synthesized taxonomy allows for a non-linear and adaptive assessment of student progress, not only at the

collective but also at the individual level. The integration of such a taxonomy within an educational platform offers the possibility of adapting the content it makes available to the student, depending on the feedback obtained by him [8].

Using this mathematical method, two or more results of the evaluation stages can be combined, and a final score can be calculated to have an accurate assessment of the level the student is at. After obtaining the weighted average, an analysis can be made on the results obtained, and the content displayed to the student further on the educational platform will be adapted so that the average is as high as possible from one lesson to another [9].

### 3 ACM Classification System

The ACM (Association for Computing Machinery) taxonomy is used in computer science as a standardized classification system. It can be used to organize and label research topics, subfields, and subject areas. Researchers, academic journals, and digital libraries that index and facilitate the search and classification of scientific content are the main actors who use it very often [10].

The ACM taxonomy had a first version in 1998. Later, the 2012 version appeared, which was developed as a poly-hierarchical ontology that can be used for semantic web applications. The previous version, which was in 1998, was used as a standard classification system for the field of computer science. It is based on a semantic vocabulary with a single source of categories and concepts and is prepared for structural changes that may be made in the future. ACM offers the possibility of using a tool within the virtual display format. This helps to apply the CCS categories to future work and ensures that it remains current and relevant. On the virtual display, users can see both interactive views and views of the classification tree [11]. The ACM taxonomy is organized so that general categories are presented first, followed by subcategories that are increasingly specific and focused on the desired topic. It presents several main

categories that are very relevant and cover major areas of computer science. These categories are as follows: (1) General and Reference, (2) Hardware, (3) Computer Systems Organizations, (4) Networks, (5) Software and its engineering, (6) Theory of Computation, (7) Mathematics of Computing, (8) Information systems, (9) Security and privacy, (10) Human-Centered computing, (11) Computing methodologies, (12) Applied Computing, (13) Social and professional topics. Using these main categories, those who want to use the information can have much easier access to it in the desired area. For example, if someone wants to access information about 3D Integrated Circuits, they must access the main category Hardware, then access the subcategory Integrated Circuits. Users can follow this example for each information need they have regarding a specific topic in the field of computer science [12]. One of the primary functions of the ACM taxonomy is to support efficient indexing and retrieval of scholarly content. When submitting an article to an ACM journal or conference, authors must select the relevant classification codes that best describe the subject matter of their work. These codes are then used by digital libraries (such as the ACM Digital Library) to categorize publications, making it easier for users to find material that is relevant to their interests [13]. Taxonomy also helps editors assign reviewers with expertise in the right fields, thereby improving the peer-review process. In addition, it is a valuable tool for students and researchers who want to identify relevant areas of research or explore existing classifications in a particular field [14].

The ACM taxonomy is a key element in standardizing the language used in computer science research. The way it is used is to promote a common vocabulary to make communication between researchers easier and to make a clearer

documentation of scientific contributions. At the same time, it offers the possibility of analyzing research trends by examining the distribution of publications as benchmarks of taxonomy codes [15]. The ACM taxonomy is very often used in the development of automatic tools that classify academic works using natural language processing (NLP) techniques. The indexing process for publishers and digital platforms is much simplified because algorithms can analyze the content of a scientific article and suggest appropriate ACM classification codes [16].

This is a tool of increasing importance in the academic and research environment of computer science. Through the coherent and hierarchical organization of the field, the indexing, analysis, and discovery of scientific content is supported. The existence of a classification system such as the ACM CCS is crucial in times of exponential data growth. These are the ability to maintain clarity and efficiency in scientific communication [17].

#### **4 NIST CSRC Taxonomy**

An essential basis for developing a structured taxonomy is the use of standardized classification systems. The taxonomies created by the National Institute of Standards and Technology (NIST) through its Computer Security Resource Center (CSRC) are one of the best-known in the field of cybersecurity [10]. The NIST CSRC taxonomy provides a comprehensive framework that covers several dimensions of cybersecurity. It contributes to making understanding, research, and development in this field easier. Furthermore, it helps identify research areas that are already well established while guiding future research directions by highlighting those that are under-explored.

There are six dimensions (topics) that form the basis of this taxonomy: (1) Security and privacy, (2) Technologies, (3) Applications, (4) Laws and regulations, (5) Activities and products, and (6) sectors [18]. The first one covers topics such as data science for

security, cybersecurity, and the social and human side of cybersecurity. It attracts input from disciplines such as management and psychology, showing the field's interdisciplinary nature [19]. The second dimension includes a wide range of technologies, among them 5G and 6G networks, and points to the need for ongoing research to address emerging vulnerabilities [20]. The Applications dimension focuses on the practical use of security measures in different sectors. It highlights the importance of context in developing solutions and encourages exploration of areas that have been little studied [21]. The laws and regulations dimension refers to the legal framework that governs cybersecurity practices. This varies by jurisdiction, and emphasizes the need for compliance and for regulations that can adapt to new threats [22]. The fifth dimension separates activities such as research, technological development, and public policy making, and underlines their dynamic character [19]. The last dimension classifies security efforts by industry, which comes with tailored approaches for recognizing that each sector faces specific challenges and rules [20].

In adaptive learning, using the NIST taxonomy provides a wide range of advantages, especially on platforms based on artificial intelligence. It allows a structured modelling of knowledge, which is the basis for building individual learner maps that can be linked to learning goals and automatically adapted to their performance and difficulties. The taxonomy also supports targeted feedback: an AI model can connect learners' errors not only to the wrong answers, but also to specific knowledge areas, for example, distinguishing between an error in the logic of a cryptographic protocol and one related to identity-federation policy. As a result, feedback becomes more precise and educationally relevant.

The NIST taxonomy also helps align the proposed AI-based STEM learning taxonomy with structures recognized worldwide. According to the report published by the Joint Research Centre (JRC) of the European Commission, titled "A Proposal for a European Cybersecurity Taxonomy" (2019), the NIST CSRC model was one of the most influential sources in shaping the European knowledge framework in the field [10].

The European taxonomy expands the NIST framework by integrating new rules. These consist of the GDPR and NIS Directive, which suit the legal and sector-specific needs of the European Union. The GDPR regulates how personal data is processed in the EU, stressing legality, transparency, and accountability [23], and applies to all entities that handle data about EU residents [24]. The NIS Directive, on the other hand, aims at the security of networks and information systems, placing security requirements and reporting duties on providers of essential and digital services [25].

The taxonomy we propose for learning in cybersecurity mixes elements from the NIST framework, the European model, and the ACM taxonomy. This aims to combine technological coverage, regulatory compliance, and sector relevance. The detailed structure of the NIST taxonomy is an ideal basis for AI learning platforms. It enables the assignment, adaptation, and dynamic assessment of educational modules. Additionally, it also makes it possible to link cognitive taxonomies (such as Bloom or Anderson) with thematic classifications in cybersecurity. Thus, learning management systems can adapt content not only to the field of expertise (for example, malware analysis) but also to the cognitive level involved (application versus evaluation).

In conclusion, the NIST CSRC taxonomy is a fundamental model for organizing adaptive learning experiences in cybersecurity. Its multidimensional structure fits the goals of personalized learning, automated feedback, and knowledge

assessment. Despite originating from an American context, its relevance and solid design make it a valuable reference for creating a global educational framework in cybersecurity.

## 5 Our proposed taxonomy

To support adaptive learning in cybersecurity education, we propose a hierarchical taxonomy integrating core elements from the NIST CSRC Taxonomy and the ACM Classification System. This structured AI-compatible model enables personalized knowledge maps aligned with STEM goals at undergraduate and master's levels.

There are three levels of taxonomy:

- Core domains
- Learning subdomains
- Cognitive learning objectives

This structure allows curriculum content to be mapped to knowledge checkpoints. As a result, it facilitates AI-driven formative and personalized feedback.

### 5.1 Core domains

The selected core domains reflect comprehensive and interdisciplinary cybersecurity challenges. The NIST CSRC and ACM Classification System serve as the foundation, addressing key technological and sector-specific issues relevant globally:

1. Cryptography
2. Identity and access management
3. Security operations and incident response
4. Network security
5. Privacy and data protection
6. Human factors and security behavior
7. Software and systems security

### 5.2 Learning subdomains

The second level consists of concise, practical subdomains designed for quick learning modules (approximately 2 hours

each). This is based on reflecting immediate workforce needs and current organizational challenges, such as:

1. Cryptography
  - a. Digital signatures and hashing
  - b. Encryption techniques
2. Identity and access management
  - a. Multi-factor authentication
  - b. Authorization and access control models
3. Security operations and incident response
  - a. Incident detection and reporting
  - b. Digital forensics basics
4. Network security
  - a. Firewalls and VPN
  - b. Wireless and mobile security
5. Privacy and data protection
  - a. GDPR compliance
  - b. Anonymization techniques
6. Human factors and security behavior
  - a. Phishing and social engineering awareness
  - b. Usability and security practices
7. Software and systems security
  - a. Secure coding practices
  - b. Vulnerability scanning and management

### 5.3 Cognitive learning objectives

At this level, each subdomain includes two clear objectives. These facilitate practical learning and highlight different skills:

- An “Apply” objective promotes immediate practical skills.
- An “Analyze/Evaluate” objective encourages complex situation analysis and critical thinking.

Examples of objectives include vulnerability analysis, configuring firewall rules, applying anonymization techniques, and evaluating GDPR compliance. These support automated assessment and personalized learning paths through AI platforms.

**Table 1.** Cognitive learning objectives in cybersecurity education

<b>Subdomain</b>	<b>Objective 1 (Apply)</b>
	<b>Objective 2 (Analyze/Evaluate)</b>
Digital signatures and hashing	Apply hashing for data integrity
	Compare SHA-256 and SHA-1 algorithms
Encryption techniques	Apply symmetric encryption for files
	Evaluate AES versus DES vulnerabilities
Multi-factor authentication	Configure MFA in real scenarios
	Analyze SMS authentication risks
Authorization and access control	Apply RBAC rules in real systems
	Compare RBAC and ABAC
Incident detection and reporting	Identify incidents using logs
	Analyze incidents via firewall logs
Digital forensics basics	Collect digital evidence
	Evaluate the validity of digital evidence
Firewalls and VPN	Configure basic firewall rules
	Analyze VPN configuration-related incidents
Wireless and mobile security	Set up secure Wi-Fi networks
	Analyze wireless protocol vulnerabilities
GDPR compliance	Identify sensitive personal data
	Evaluate GDPR compliance in web forms
Anonymization techniques	Apply anonymization methods
	Compare anonymization and pseudonymization
Phishing and social engineering	Identify phishing emails
	Analyze the effectiveness of anti-phishing campaigns
Usability and security practices	Implement basic security controls
	Analyze usability impacts on security
Secure coding practices	Apply OWASP principles to application code
	Analyze code for OWASP Top 10 vulnerabilities
Vulnerability scanning and management	Conduct vulnerability scanning
	Evaluate the identified vulnerabilities' criticality

The proposed taxonomy integrates seamlessly with AI-driven adaptive learning platforms. Each taxonomy element serves as a personalized learning checkpoint. This allows immediate and

targeted feedback based on identified gaps. Thus, the taxonomy becomes a robust tool for rapid and personalized learning improvement.



## 6 Conclusion

This study proposes an integrative taxonomy designed to support adaptive learning in cybersecurity education, addressing a critical need for personalized, scalable, and cognitively aligned instructional systems. By unifying elements from the NIST CSRC taxonomy, the ACM Classification System, and cognitive learning models, we construct a robust framework capable of guiding both the development of content and the automation of formative assessment within AI-based educational platforms.

The taxonomy's hierarchical structure - comprising core domains, focused subdomains, and cognitive learning objectives - enables the construction of individualized learning maps that respond dynamically to student performance. Through this structure, educators and systems alike can more effectively deliver targeted feedback, monitor conceptual understanding, and adjust instructional strategies in real time. Importantly, the taxonomy supports alignment with regulatory standards such as the GDPR and the NIS Directive, ensuring that the educational content is not only pedagogically sound but also compliant with legal and ethical requirements relevant to cybersecurity. Future research should focus on empirical validation of this taxonomy through real-world deployment in university-level courses or professional training programs. Additionally, further iterations may explore expansion into related domains such as data science, digital ethics, or critical infrastructure protection. Through these next steps, the taxonomy can evolve from a conceptual proposal to a scalable standard for adaptive learning in cybersecurity and beyond.

## References

- [1] D. M. Djamalovna, "METHODS AND TOOLS FOR ASSESSING

STUDENT COMPETENCIES," *Current Research Journal of Philological Sciences*, vol. 5, no. 10, pp. 19–24, Oct. 2024, doi: 10.37547/PHILOLOGICAL-CRJPS-05-10-04.

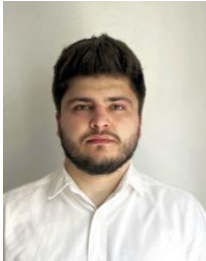
- [2] A. Kinder et al., "Effects of adaptive feedback generated by a large language model: A case study in teacher education," *Computers and Education: Artificial Intelligence*, vol. 8, p. 100349, Jun. 2025, doi: 10.1016/J.CAEAI.2024.100349.
- [3] S. Huskisson, T. O'Mahony, and S. Lacey, "Improving student outcomes using automated feedback in a first-year economics class," *International Review of Economics Education*, vol. 47, p. 100303, Dec. 2024, doi: 10.1016/J.IREE.2024.100303.
- [4] M. Cooper-Stachowsky and M. Cooper-Stachowsky, "Enhancing Learning via AI-Generated Feedback and Resubmission of Formative Assessments," *Proceedings of the Canadian Engineering Education Association (CEEAA)*, Dec. 2024, doi: 10.24908/pceea.2024.18540.
- [5] H. E. Sari, B. Tumanggor, and D. Efron, "Improving Educational Outcomes Through Adaptive Learning Systems using AI," *International Transactions on Artificial Intelligence*, vol. 3, no. 1, pp. 21–31, Nov. 2024, doi: 10.33050/ITALIC.V3I1.647.
- [6] E. du Plooy, D. Casteleijn, and D. Franzsen, "Personalized adaptive learning in higher education: A scoping review of key characteristics and impact on academic performance and engagement," *Heliyon*, vol. 10, no. 21, p. e39630, Nov. 2024, doi: 10.1016/J.HELİYON.2024.E39630.
- [7] M. A. Aripin, R. Hamzah, P. Setya, M. H. M. Hisham, and M. I. Mohd Ishar, "Unveiling a new taxonomy in education field," *International Journal of Evaluation and Research in Education (IJERE)*, vol. 9, no. 3, pp. 524–530, Aug. 2020, doi: 10.11591/IJERE.V9I3.20458.

- [8] O. Gun and M. J. Bossé, "Synthesizing cognitive mathematics learning taxonomies," *Thinking Skills and Creativity*, vol. 57, p. 101796, Sep. 2025, doi: 10.1016/J.TSC.2025.101796.
- [9] S. Sarker, M. K. Paul, S. T. H. Thasin, and M. A. M. Hasan, "Analyzing students' academic performance using educational data mining," *Computers and Education: Artificial Intelligence*, vol. 7, p. 100263, Dec. 2024, doi: 10.1016/J.CAEAI.2024.100263.
- [10] I. Nai-Fovino, R. Neisse, J. L. Hernandez-Ramos, N. Polemi, G. Ruzzante, and M. Figwer, "A Proposal for a European Cybersecurity Taxonomy," 2019, doi: 10.2760/106002.
- [11] "Computing Classification System." Accessed: May 30, 2025. [Online]. Available: <https://dl.acm.org/ccs>
- [12] "HOW TO CLASSIFY WORKS USING ACM'S COMPUTING CLASSIFICATION SYSTEM".
- [13] S. Isukapalli and S. N. Srirama, "A systematic survey on fault-tolerant solutions for distributed data analytics: Taxonomy, comparison, and future directions," *Computer Science Review*, vol. 53, p. 100660, Aug. 2024, doi: 10.1016/J.COSREV.2024.100660.
- [14] D. Carreira-Flores, M. Rubal, E. Cabecinha, G. Díaz-Agras, and P. T. Gomes, "Unveiling the role of taxonomic sufficiency for enhanced ecosystem monitoring," *Marine Environmental Research*, vol. 200, p. 106631, Sep. 2024, doi: 10.1016/J.MARENVRES.2024.106631.
- [15] J. B. Minani et al., "IoT systems testing: Taxonomy, empirical findings, and recommendations," *Journal of Systems and Software*, vol. 226, p. 112408, Aug. 2025, doi: 10.1016/J.JSS.2025.112408.
- [16] M. P. dos Santos, W. F. Heckler, R. S. Bavaresco, and J. L. V. Barbosa, "Machine learning applied to digital phenotyping: A systematic literature review and taxonomy," *Computers in Human Behavior*, vol. 161, p. 108422, Dec. 2024, doi: 10.1016/J.CHB.2024.108422.
- [17] R. L. Fleurence et al., "A Taxonomy of Generative AI in HEOR: Concepts, Emerging Applications, and Advanced Tools – An ISPOR Working Group Report," *Value in Health*, vol. 0, no. 0, May 2025, doi: 10.1016/J.JVAL.2025.04.2167.
- [18] "NIST Computer Security Resource Center | CSRC." Accessed: May 30, 2025. [Online]. Available: <https://csrc.nist.gov/>
- [19] H. Suryotrisongko and Y. Musashi, "Review of cybersecurity research topics, taxonomy and challenges: Interdisciplinary perspective," *Proceedings - 2019 IEEE 12th Conference on Service-Oriented Computing and Applications, SOCA 2019*, pp. 162–167, Nov. 2019, doi: 10.1109/SOCA.2019.00031.
- [20] J. E. Ounza, "A taxonomical survey of 5G and 6G security and privacy issues," <https://gjeta.com/sites/default/files/GJET-A-2023-0047.pdf>, vol. 14, no. 3, pp. 042–060, Mar. 2023, doi: 10.30574/GJETA.2023.14.3.0047.
- [21] C. Meadows, "An outline of a taxonomy of computer security research and development," *Proceedings New Security Paradigms Workshop*, vol. Part F129673, pp. 33–35, Aug. 1993, doi: 10.1145/283751.283770/ASSET/8949DF88-571F-4546-B40C-C23C8C7007E4/ASSETS/283751.283770.FP.PNG.
- [22] G. M. NIST, "NIST Privacy Workforce Taxonomy," 2024. doi: 10.6028/NIST.CSWP.38.ipd.
- [23] Andrew Cormack, "An Introduction to the GDPR (v2)," vol. 1, no. 5, 2021.
- [24] "General Data Protection Regulation," A Guide to Financial Regulation for

Fintech Entrepreneurs, pp. 187–194, Mar. 2018, doi: 10.1002/9781119436775.CH18.

[25] M. D. Cole and S. Schmitz, “The

Interplay Between the NIS Directive and the GDPR in a Cybersecurity Threat Landscape,” SSRN Electronic Journal, Dec. 2019, doi: 10.2139/SSRN.3512093.



**Mihai-Nicolae DULGHERU** completed his studies at the Faculty of Cybernetics, Statistics and Economic Informatics, Bucharest University of Economic Studies, in 2022. He then earned a master’s degree in eBusiness at the same institution and is currently a first-year doctoral student at the Doctoral School of Economic Informatics, where he also serves as an associate teaching assistant in the Department of Economic Informatics and Cybernetics. His doctoral research focuses on **Artificial-Intelligence-based adaptive learning technologies**, with a thesis titled “Design and development of AI-driven adaptive learning solutions”. Beyond academia, he has three years of professional experience as a full-stack web developer, possessing solid expertise in object-oriented and functional programming, modern web frameworks, distributed systems, and cloud-native architectures. His broader research and teaching interests include educational technology, intelligent information systems, and software engineering.



**Andrei-Cătălin NICA** graduated from the master’s program in Online Marketing at the Bucharest University of Economic Studies in 2024. He has been a PhD student since 2024 within the Doctoral School of Economic Informatics at the Bucharest University of Economic Studies, currently enrolled in the second year. He has conducted Economic Informatics seminars for seven student groups from the Faculties of Business and Tourism, and International Economic Relations. At present, he works as a programmer in a company specializing in robotics equipment and software development in the educational field. He has professional experience in the fields of STEM, artificial intelligence, and educational robotics. Among his works is the article: “Supporting Hybrid and Remote Learning through Virtual Infrastructures Integrated with Cyberquest and Raspberry Pi” (2025).



**Cătălin-Alexandru TRANDAFIRIDIS** graduated from the Faculty of Transport Engineering, University Politehnica of Bucharest, with a specialization in Electronics, Remote Control and Telecommunications. He pursued a master’s degree in Intelligent Transport Systems within the same institution, focusing on the integration of cybersecurity measures in modern transportation infrastructures. Professionally, he has contributed to digital infrastructure and ITS-related projects in collaboration with the Romanian Ministry of Transport and European initiatives such as DATEX and NAPCORE. He is the founder and managing partner of a technology and digital marketing company and teaches computer science at the high school level within an international academic environment. His work focuses on applied cybersecurity, system architecture, and educational technologies.